

Variable Selection in High Dimensions with Random Designs and Orthogonal Matching Pursuit

Antony Joseph*

August 29, 2011

Abstract

The performance of Orthogonal Matching Pursuit (OMP) for variable selection is analyzed for random designs. When contrasted with the deterministic case, since the performance is here measured after averaging over the distribution of the design matrix, one can have far less stringent sparsity constraints on the coefficient vector. We demonstrate that for exact sparse vectors, the performance of the OMP is similar to known results on the Lasso algorithm [*IEEE Trans. Inform. Theory* **55** (2009) 2183-2202]. Moreover, variable selection under a more relaxed sparsity assumption on the coefficient vector, whereby one has only control on the ℓ_1 norm of the smaller coefficients, is also analyzed. As a consequence of these results, we also show that the coefficient estimate satisfies strong oracle type inequalities.

1 Introduction

Consider linear regression model,

$$Y = X\beta + \epsilon \tag{1}$$

where $X \in \mathbb{R}^{n \times p}$, the coefficient vector $\beta \in \mathbb{R}^p$ and noise $\epsilon \in \mathbb{R}^n$. The high dimensional case, where p is of the same order, or possibly much larger than n , has been of immense interest nowadays. In many applications, interest is not primarily on prediction of the response Y , but on the accuracy of estimation of the coefficient β . Examples of such applications include, micro-array data analysis, graphical model selection [19], compressed sensing [10], [9], and in communications [3],[2],[24]. As is well known, in the high dimensional setting, β is unidentifiable unless the design matrix X is well-structured and there is some sparsity constraint

*Department of Statistics, Yale University , New Haven, CT 06520 USA, e-mail : antony.joseph@yale.edu

on the coefficient vector β . This sparsity assumption corresponds to restricting β to few non-zero entries (ℓ_0 -sparsity), or more generally, assuming that β has only few terms that are large in magnitude.

The Orthogonal Matching Pursuit [20] is a variant of the Matching Pursuit algorithm [18], where, successive fits are computed through the least squares projection of Y on the current set of selected terms. For deterministic X matrices, variable selection properties of this algorithm, for ℓ_0 -sparse vectors, have been analyzed for the noisy case in Zhang [28] and Cai and Wang [5]. However, as we shall review Subsection 1.2, although they give strong performance guarantees under certain conditions on the X matrix, they impose severe constraints on the sparsity of β . Similar results have been shown for the Lasso, for example in Zhao and Yu [30]. With random designs one can have reliable detection of the support with far less stringent sparsity constraints; the performance is here measured after averaging over the distribution of X . For example, Wainwright [26] proved such results for the Lasso algorithm. The main results of this paper, apart from showing that similar properties hold for the OMP, demonstrate two important additional properties. Firstly, we give results on partial support recovery, which is important since exact recovery of support places strong requirements on n if some of the non-zero elements are small in magnitude. Secondly, and more importantly, we relax the assumption that β is ℓ_0 -sparse and address variable selection under a more general notion of sparsity, whereby one has only control on the ℓ_1 norm of the smaller elements of β . We demonstrate that even under this more relaxed assumption, one can reliably estimate the position of the larger entries using the OMP. This has certain parallels with recent work on the Lasso by Zhang and Huang [27]. As a consequence of these results, we show that our coefficient estimate, after running the algorithm, satisfies strong oracle inequalities, similar to that demonstrated for the Lasso [29] and Dantzig selector [6].

The paper is organized as follows. Below, we describe the OMP algorithm. The stopping criterion we use is slightly different from what is traditionally used in literature. Subsection 1.2 motivates in greater detail our interest in random designs. In Subsection 2.1 we give results for design matrices that have i.i.d sub-Gaussian entries and ℓ_0 -sparse vectors. This extends the results in Tropp and Gilbert [25] for the noisy case. In Subsection 2.2 we describe more general results with correlated Gaussian designs, where we only have control over the ℓ_1 norm of the smaller coefficients. Sections 3, 4 and 5 gives proofs of our main results. The appendices contains auxiliary results.

1.1 The Orthogonal Matching Pursuit algorithm

Denote as $J = J_1 = \{1, 2, \dots, p\}$ to be the set of indices corresponding to columns in the X matrix. For each step i , with $i \geq 1$, a single index $a(i)$ is detected to be non-zero in that step. Accordingly, denoting $d(i) = a(1) \cup a(2) \dots \cup a(i)$ as the set of detected columns after i steps, step $i + 1$ of the algorithm only operates on the columns in $J_{i+1} = J - d(i)$, that is, the columns not detected in the previous steps. In other

words, indices detected in previous steps remain detected.

The decision on whether a particular index j is detected during a particular step i is based on the absolute value of a statistic \mathcal{Z}_{ij} . Here, \mathcal{Z}_{ij} is simply the inner product between X_j and the normalized residual R_{i-1} computed for the previous step.

Apart from the response vector Y and design matrix X , the other input to the algorithm is a positive threshold value τ . Denote $\|\cdot\|$ as the euclidean norm. We now describe the OMP algorithm.

- Initialize $R_0 = Y$, $d(0) = \emptyset$. Start with step $i = 1$.

- Update

$$\mathcal{Z}_{ij} = X_j^T \frac{R_{i-1}}{\|R_{i-1}\|}, \quad \text{for } j \in J_i.$$

- If $\max_{j \in J_i} |\mathcal{Z}_{ij}| > \tau$, do the following:
 - Assign $a(i) = \arg \max\{|\mathcal{Z}_{i,j}| : j \in J_i\}$.
 - Set $d(i) = d(i-1) \cup a(i)$. Update $R_i = (I - \mathcal{P}_i)Y$, where \mathcal{P}_i is the projection matrix for the column space of $X_{d(i)}$, and set $J_{i+1} = J_i - a(i)$.
 - Increase i by one and go to step 2.
- Stop if $\max_{j \in J_i} |\mathcal{Z}_{ij}| \leq \tau$.

We remark that for any step i , the inner product $X_j^T R_{i-1}$, for $j \in d(i-1)$, is 0. Correspondingly, since $\mathcal{Z}_{ij} = 0$, for $j \in d(i-1)$, the maximum of \mathcal{Z}_{ij} over $j \in J_i$, is the same as the maximum over all $j \in J$. Also, the newly selected term $a(i)$ may be equivalently expressed as,

$$a(i) = \arg \min_{j \in J} \inf_{w \in \mathbb{R}} \|Y - \text{Fit}_{i-1} - wX_j\|^2,$$

where Fit_{i-1} is the least squares fit of Y on the columns in $d(i-1)$. In this respect, the OMP is similar to other greedy algorithms such as relaxed greedy and forward-stepwise algorithms ([4], [15], [16], [17]), that operate through successive reduction in the approximation error.

As mentioned earlier, the stopping criterion considered here is slightly different from that considered in literature. Traditionally, for the no noise setting, the algorithm is run until there is a perfect fit between Y and the selected terms, that is $R_i = 0$ (see for example [23], [25]). In the noisy case, as analyzed over here, there are two standard approaches. The first, as done in [5], [28], is to stop when $\max_{j \in J} |X_j^T R_{i-1}|$ is less than some fixed threshold. The second approach, as analyzed in [12], [5], is to stop when $\|R_i\|$ is less than some pre-specified value.

Our stopping criterion, which is more similar to the first approach, is equivalent to continuing the algorithm until $\max_{j \in J} |X_j^T R_{i-1}| \leq \tau \|R_{i-1}\|$. The motivation for the use of such a statistic comes from the analysis of a similar iterative algorithm in Barron and Joseph [2] for a communications setting. However, there the values of the non-zero β_j 's were known in advance; this added information played an important role in the analysis of the algorithm. A similar statistic was used by Fletcher and Rangan [14] for an asymptotic analysis of the OMP for exact support recovery using i.i.d designs.

Notation: Let $a = a(n, p, k)$, $b = b(n, p, k)$ be two positive functions of n, p and k . We denote as $a = O(b)$, if $a \leq c_1 b$ for some constant positive constant c_1 that is independent of n, p or k . Similarly, $a = \Omega(b)$ means $a \geq c_2 b$ for positive c_2 independent of n, p or k .

1.2 Related work

As mentioned earlier, we are interested in variable selection in the high dimensional setting. Apart from iterative schemes, another popular approach is the convex relaxation scheme Lasso [22]. In order to motivate our interest in random design matrices, we describe existing results on variable selection, using both methods, with deterministic as well as random design matrices. For convenience, we concentrate on implications of these results assuming the simplest sparsity constraint on β , namely that β has only a few non-zero entries.

In particular, we assume that,

$$|S_0(\beta)| = k, \quad \text{where } S_0(\beta) = \{j : \beta_j \neq 0\}. \quad (2)$$

In other words, attention is restricted to all k -sparse vectors, that is, those that have exactly k non-zero entries. For convenience, we drop the dependence on β and denote $S_0(\beta)$ as S_0 whenever there is no ambiguity. The simplest goal then is to recover S_0 exactly, under the additional assumption that all β_j , for $j \in S_0$, have magnitude at least β_{min} , where $\beta_{min} > 0$. Denote as $\mathcal{C} \equiv \mathcal{C}(\beta_{min}, k)$, as the set of coefficient vectors satisfying this assumption.

Further, denote \hat{S} as the estimate of S_0 obtained using either method, and $\mathcal{E} = \{\hat{S} \neq S_0\}$ the error event that one is not able to recover the support exactly. For deterministic X , interest is mainly on conditions on X so that

$$P_{err, X} = \sup_{\beta \in \mathcal{C}} \mathbb{P}_\beta(\mathcal{E} | X) \quad (3)$$

can be made arbitrarily small when n, p , or k become large. Here $\mathbb{P}_\beta(\cdot | X)$ denotes the distribution of Y for the given X and β .

A common sufficient condition on X for this type of recovery is the *mutual incoherence condition*, which requires that the inner product between distinct columns be small. In particular, letting $\|X_j\|^2/n = 1$,

for all $j \in J$, it is assumed that

$$\gamma(X) = \frac{1}{n} \max_{j \neq j'} |X_j^T X_{j'}| \quad (4)$$

is $O(1/k)$. Another related criterion is the *irrepresentable criterion* [23], [30], which assumes, for all subset T of size k , that

$$\|(X_T^T X_T)^{-1} X_T^T X_j\|_1 < 1, \quad \text{for all } j \in J - T. \quad (5)$$

Here $\|\cdot\|_1$ denotes the ℓ_1 norm.

Observe that if $P_{err,X}$ (3) is small, it gives strong guarantees on support recovery, since it ensures that any β , with $|S_0(\beta)| = k$, can be recovered with high probability. However, it imposes severe constraints on the X matrix. As an example, when the entries of X are i.i.d Gaussian, the coherence $\gamma(X)$ is around $\sqrt{2 \log p / n}$. Correspondingly, for (4) to hold, n needs to be $\Omega(k^2 \log p)$. In other words, the sparsity k should be $O(\sqrt{n / \log p})$, which is rather strong since ideally one would like k to be of the same order as n . Similar requirements are needed for the irrepresentable condition to hold. Recovery using the irrepresentable condition has been shown for Lasso in [30], [26], and for the OMP in [28], [5]. Indeed, it has been observed, in [30] for the Lasso, and in [28], for the OMP, that a similar such condition is also necessary if one wanted exact recovery of the support, while keeping $P_{err,X}$ small.

A natural question is to ask about requirements on X to ensure recovery in an average sense, as opposed to the strong sense described above. One way to proceed, as done over here, is to consider random X matrices and ask about the requirements on n , p , k , as well as β_{min} , so that

$$P_{err} = \sup_{\beta \in \mathcal{C}} \mathbb{P}_\beta(\mathcal{E}) \quad (6)$$

is small. Here $\mathbb{P}_\beta(\mathcal{E}) = E_X \mathbb{P}_\beta(\mathcal{E}|X)$, where the expectation on the right is over the distribution of X . For the Lasso, Wainwright [26] considers random X matrices, with rows drawn i.i.d $N_p(0, \Sigma)$. It is shown that under certain conditions on Σ , which can be described as population counterparts of the conditions for deterministic X 's, one can recover S_0 with high probability with $n = \Omega(k \log p)$ observations, with the constant depending inversely on β_{min}^2 . The form of n is in a sense ideal since now $k = O(n / \log p)$ is nearly the same n , if we ignore the $\log p$ factor. As mentioned earlier, apart from establishing similar properties to hold for the OMP with k -sparse vectors, we also demonstrate strong support recovery results under a more general notion of sparsity. These results are described in the next section.

We also note that instead of averaging over X , one could assume a distribution on β and analyze the average probability of \mathcal{E} over this distribution. This is done in Candès and Plan [7] for the Lasso. Here, for fixed magnitudes of the k non-zero β , the support of β is uniformly assigned over all possible subsets of size k . Once the support is chosen, the signs for the non-zero β_j 's are assigned ± 1 with equal probability. If $\text{Avg}[\cdot]$ denotes the expectation with this distribution of β , it is shown that one could keep $\text{Avg}[\mathbb{P}_\beta(\mathcal{E}|X)]$ low for $\gamma(X)$ as high as $O(1 / \log p)$. This condition on $\gamma(X)$ is less stringent than before and leads to a demonstration

that $n = \Omega(k \log p)$ is sufficient for support recovery, provided $\|X\| \approx \sqrt{p/n}$, where $\|\cdot\|$ denotes the spectral norm. We provide comparisons with this work in Section 6.

Notation: For a set $\mathcal{A} \subseteq J$, we denote as $X_{\mathcal{A}}$ the sub-matrix of X comprising of columns with indices in \mathcal{A} . Similarly, for any $p \times 1$ vector β , we denote as $\beta_{\mathcal{A}}$ the $|\mathcal{A}| \times 1$ sub-vector with indices in \mathcal{A} . Also let $\mathcal{A}^c = J - \mathcal{A}$.

2 Results

Before discussing our main results with Gaussian matrices, in Subsection 2.1 we state results when the entries of X are i.i.d sub-Gaussian and when the vector β has k non-zero entries. The noise vector is also assumed to come from a sub-Gaussian distribution with scale σ . This generalizes the results of Tropp and Gilbert [25] for the noisy case. While preparing this manuscript we discovered that Fletcher and Rangan [14] have analyzed the OMP for i.i.d designs and for k -sparse vectors, similar to that in Subsection 2.1. However, there the analysis was for exact support recovery and was asymptotic in nature. Further, they focused on a specific regime, where $k\beta_{\min}^2/\sigma^2$ tends to infinity. We provide more comparisons with this work later on in the paper.

We show that $n = \Omega(k \log p)$ samples are sufficient for the recovery of any coefficient vector with β_{\min} that is at least the same order as the *noise level*. More specifically, define

$$\mu_n = \sqrt{(2 \log p)/n}. \quad (7)$$

The quantity $\sigma\mu_n$ can be thought of as the noise level. To see why this is so, consider the orthogonal design where $X^T X/n = I$ and noise $\epsilon \sim N(0, \sigma^2 I)$. Assume that, as usual, we are interested in recovering any β with $|S_0(\beta)| = k$. A natural estimate of the support would be,

$$\hat{S} = \{j : |z_j| > t\} \quad \text{with} \quad z_j = X_j^T Y/n, \quad (8)$$

where t is positive. Notice that $z_j \sim N(\beta_j, \sigma^2/n)$ for each $j \in J$. Correspondingly, since $z_j \sim N(0, \sigma^2/n)$, for $j \in J - S_0$, one sees that t has to be of the form $\sigma\mu_n$ in order to prevent false discoveries with high probability. Similarly β_j , for all $j \in S_0$, has to have magnitude at least $\sigma\mu_n$ if one wanted to avoid false negatives.

The analysis of iid designs, as done in Subsection 2.1, forms an important ingredient to compressed sensing [9], [10]. However, it may not be useful for statistical applications, where typically the choice of the X matrix is not under one's control. Accordingly, in Subsection 2.2, we assume that the rows of X are drawn i.i.d from $N_p(0, \Sigma)$, with certain assumptions on Σ . This model was also employed to detect the neighborhood of a

node in high dimensional graphs by Meinshausen and Bühlmann [19]. Moreover, we relax the assumption that β is k -sparse and only assume that there is a set $S = S(\beta)$, of size k , such that β_{S^c} is sparse in a more general sense. Here β_{S^c} denotes the vector of coefficients outside of S . More specifically, for a constant $\nu \geq 0$, if

$$S = \{j : |\beta_j| > \sigma\nu\mu_n\}, \quad \text{with } |S| = k, \quad (9)$$

we assume

$$\|\beta_{S^c}\|_1 \leq \sigma\eta\mu_n, \quad (10)$$

for an appropriately chosen η . A natural choice would be to take $\nu = 1$. Then, S would correspond to the indices above the noise level. We show that for η not too large, the OMP can detect the large indices in S with high probability, provided Σ satisfies certain conditions. As a consequence of these results, we show that the coefficient estimate satisfies strong oracle inequalities.

2.1 Recovery with sub-Gaussian designs

In this section we address the requirements on n, p, k as well as β_{min} , to recover the support of β , either exactly or nearly so, where we assume that $|S_0(\beta)| = k$. Here $S_0(\beta)$ is as in (2). We allow the case that k may be zero. Further, since it may not be a realistic assumption that k is known, we assume that we only know an upper bound \bar{k} on k , with $\bar{k} \geq \max\{k, 1\}$.

Let $X_{\ell j}$, for $\ell = 1, \dots, n$ and $j = 1, \dots, p$, denote the entries of the X matrix. Throughout this section we assume that the $X_{\ell j}$'s are independent sub-Gaussian with mean 0 and scale 1, that is $\mathbb{E}e^{tX_{\ell j}} \leq e^{t^2/2}$, for $t \in \mathbb{R}$. Further, we assume that the noise vector ϵ is independent of X and has independent sub-gaussian entries with mean 0 and scale σ , that is $\mathbb{E}e^{t\epsilon_\ell} \leq e^{\sigma^2 t^2/2}$, for $t \in \mathbb{R}$, $\ell = 1, \dots, n$. Additionally, if $k \geq 1$, we assume that the following two conditions are satisfied with high probability.

Condition 1. There exists $\lambda_{max} \geq \lambda_{min} > 0$, so that the eigenvalues of $X_{S_0}^T X_{S_0}/n$ are between λ_{min} and λ_{max} , that is

$$\lambda_{max}\|v\|^2 \geq \|X_{S_0}v\|^2/n \geq \lambda_{min}\|v\|^2 \quad \text{for all } v \in \mathbb{R}^k.$$

Condition 2. The ℓ_2 norm of the noise vector is bounded, that is $\|\epsilon\|^2/n \leq \sigma^2\lambda$, for some $\lambda > 0$.

Let \mathcal{E}_{cond} be the event that Conditions 1 or 2 fail. The first assumption is related to the restricted isometry property (Candes and Tao [8]) and the sparse eigenvalues conditions (Zhang and Huang [27]). Condition 1 is satisfied for a wide variety of random ensembles. For example, it is satisfied with high probability for the Gaussian ensemble, where the $X_{\ell j}$ are i.i.d $N(0, 1)$ and the binary ensemble, where the $X_{\ell j}$ are i.i.d uniform on $\{-1, +1\}$ (see for example, Baraniuk et al. [1]). Notice that since we are interested in controlling the

probability P_{err} in (6), because of the averaging over X , we do require that the Condition 1 hold uniformly over all S_0 , with $|S_0| = k$. Condition 2, which bounds the ℓ_2 norm of the noise vector, is required for controlling the norm of the residuals R_i . It is satisfied with high probability, for example, when the noise $\epsilon \sim N(0, \sigma^2)$.

Below, we state the theorem giving sufficient conditions on n for reliable recovery of the support of β . The threshold τ is taken to be

$$\tau = \sqrt{2(1+a)\log p}, \quad (11)$$

for some $a > 0$. Here n will be a function \bar{k} and p , as well as the various quantities defined above. The results of course hold with \bar{k} replaced by k , provided k is non-zero. In particular, for $\alpha, \delta > 0$, define

$$\xi \equiv \xi(\alpha, \delta) = \max \left\{ (1+\delta)r_1, \sigma^2 r_2^2 f(\delta)/(\bar{k}\alpha) \right\}. \quad (12)$$

where,

$$r_1 = \frac{\max\{\lambda_{max}, \lambda\}}{\lambda_{min}^3} \quad r_2 = \left[\frac{1}{\sqrt{\lambda_{min}}} + \sqrt{r_1} \right] \quad (13)$$

and

$$f(\delta) = \frac{1}{(1 - 1/\sqrt{1+\delta})^2} \quad (14)$$

Denote as $\hat{S} = \hat{S}(Y, X, \tau)$, the estimate of the support obtained after running the algorithm with the given Y, X and threshold τ . Further, denote the undetected elements of the support as $\hat{F} = S_0 - \hat{S}$. The theorem below, provides bounds on $\sum_{j \in \hat{F}} \beta_j^2$, the signal strength of the undetected components; here we assume that

$$\sum_{j \in \hat{F}} \beta_j^2 = 0 \text{ if } \hat{F} = \emptyset.$$

The following function of k characterizes the probability of failure of the algorithm.

$$p_{err, k} = \mathbb{P}(\mathcal{E}_{cond}) + 2(k+1)/p^a + 2k/p^{1+a}, \quad \text{for } k \geq 1, \quad (15)$$

and $p_{err, 0} = 2/p^a$. Here, recall that \mathcal{E}_{cond} is event that Conditions 1 or 2 fail. Notice that $p_{err, k} \leq p_{err, \bar{k}}$, since $k \leq \bar{k}$.

Regarding the choice of a , if k is $O(\log p)$, then a can be taken to be slightly larger than 0 for $p_{err, k}$ to be small, assuming p is large; however, if k scales, for example, linearly with p , then a needs to be taken to be larger than 1. We now state our theorem.

Theorem 2.1. *Let the threshold τ be as in (11). Further, let n be of the form*

$$n = \xi \bar{k} \tau^2, \quad (16)$$

with ξ as in (12).

Then, if $k \geq 1$, the following condition holds, except on a set with probability $p_{err,k}$:

$$\hat{S} \subseteq S_0 \quad \text{and} \quad \sum_{j \in \hat{F}} \beta_j^2 \leq \alpha |\hat{F}|. \quad (17)$$

In particular, if $\beta_{min}^2 > \alpha$ then $\hat{S} = S_0$, that is the support is recovered exactly, with probability at least $1 - p_{err,k}$.

If $k = 0$, $\hat{S} = \emptyset$ with probability at least $1 - p_{err,0}$.

Notice that α controls accuracy to which the support is estimated. Assuming \hat{F} is non-empty, another way of stating the theorem is that the average signal strength of the undetected components, that is $\|\beta_{\hat{F}}\|^2/|\hat{F}|$, is at most α . It may seem desirable to make α as small as possible, however, doing so increases the value of n in (16), since n is inversely related to α through $\xi(\alpha, \delta)$. Further, if α is taken to be less than β_{min}^2 , then the above theorem guarantees exact recovery of the support. Correspondingly, from (16) and (12), one sees that if

$$n = \max \left\{ b_1 \bar{k}, \frac{b_2}{\beta_{min}^2} \right\} \log p,$$

for some $b_1, b_2 > 0$, then the support can be recovered exactly with high probability.

The following corollary, which is a consequence of Theorem 2.4, shows that if $n = \Omega(\bar{k} \log p)$, one can reliably detect the indices with large coefficient values, while ensuring that there are no false discoveries. Further, if all the non zero components are above the noise level (up to a constant factor), one can estimate the support exactly with the same number of observations.

Corollary 2.2. Define $\bar{\xi} = 32r_2^2(1+a)$ and $r = 2r_2\sqrt{1+a}$. Let

$$n \geq \bar{\xi} \bar{k} \log p. \quad (18)$$

Then, if $k \geq 1$, with probability at least $1 - p_{err,k}$, the estimate \hat{S} is contained in S_0 and further,

$$\left\{ j : |\beta_j| > r \sigma \sqrt{k} \mu_n \right\} \subseteq \hat{S}.$$

Further, if $\beta_{min} > r \sigma \mu_n$, then algorithm can recover the entire support of β , that is $\hat{S} = S_0$, with probability at least $1 - p_{err,k}$.

If $k = 0$, then $\hat{S} = \emptyset$ with probability at least $1 - p_{err,0}$. Here $p_{err,\cdot}$ is as in (15).

2.2 More general results with Gaussian designs

For Gaussian ensembles, the methods used in the proof of Theorem 2.1 can be extended to give more general results on support recovery. In particular, we relax the assumption that X has i.i.d entries and assume that

rows of the X matrix are i.i.d $N_p(0, \Sigma)$. The noise vector is assumed to be independent of X , with entries i.i.d. $N(0, \sigma^2)$. As mentioned earlier, here we also address a more general type of variable selection question, where we are not interested in recovering all non-zero entries but only the ones that are large compared to the noise level. In particular, for a constant $\nu \geq 0$, let S be a set of size k as in (9), consisting of the indices corresponding to the larger elements (in magnitude) of β . Once again, we do not assume that k is known, but only assume that we have an upper bound \bar{k} on k , with $\bar{k} \geq 1$. Unlike before, we do not require that the coefficients outside of S are zero, but only assume that $\|\beta_{S^c}\|_1 \leq \sigma\eta\mu_n$, where η is allowed to scale at most linearly with \bar{k} , that is we assume that $\bar{\eta} = \eta/\bar{k}$ is $O(1)$.

Through a permutation of the columns one can, without loss of generality, write Σ as

$$\Sigma = \begin{bmatrix} \Sigma_{SS} & \Sigma_{SS^c} \\ \Sigma_{S^cS} & \Sigma_{S^cS^c} \end{bmatrix},$$

where for $\mathcal{A}, \mathcal{A}' \subseteq J$, $\Sigma_{\mathcal{A}, \mathcal{A}'} = \text{Cov}(X_{1,\mathcal{A}}, X_{1,\mathcal{A}'})$ is the covariance matrix between terms in \mathcal{A} and \mathcal{A}' . We denote the elements of the matrix as σ_{ij} , or Σ_{ij} , and use both notations interchangeably. Without loss, we assume that $\sigma_{jj} = 1$ for all j , since if this were not the case, we could always scale the coefficient vector to produce such a correlation matrix.

We make the following assumptions on the correlation matrix Σ , when $k \geq 1$. These are essentially population analogs of the sparse eigenvalue and the irrerepresentable conditions respectively.

1. There exists $s_{min}, s_{max} > 0$ so that,

$$\lambda_{min}(\Sigma_{TT}) \geq s_{min} \quad \text{and} \quad \lambda_{max}(\Sigma_{TT}) \leq s_{max}, \quad (19)$$

uniformly for all subsets T , with $|T| = k$. Here $\lambda_{min}(A)$, $\lambda_{max}(A)$ denotes the minimum and maximum eigenvalues respectively of a square matrix A .

2. For some $\omega \in [0, 1)$, the following holds,

$$\max_{j \in J-T} \|\Sigma_{TT}^{-1} \Sigma_{Tj}\|_1 \leq \omega, \quad (20)$$

uniformly for all subsets T of size k . This is essentially the population analog of the irrerepresentable condition (5).

Additionally, for $k \geq 1$, we make the following assumption that imposes bounds on certain interactions between β_{S^c} and the correlation matrix Σ . As stated below, they are not very intuitive. Lemma 2.3, however, shows that under a simple condition, which controls the magnitude of correlations of the off diagonal elements of Σ , and along with (10), one can show (19) - (21) to hold.

Let $\Sigma_{S^c|S} = \Sigma_{S^c S^c} - \Sigma_{S^c S} \Sigma_{SS}^{-1} \Sigma_{SS^c}$, denote the variance of the conditional distribution of X_{1,S^c} given $X_{1,S}$, where we recall that S is the subset of indices comprising of the k largest elements (in magnitude) of β . Let μ_n be as in (7). We make the following additional assumption.

3. For constants $\nu_1, \tilde{\nu}_1 \geq 0$, the following holds,

$$\|\Sigma_{SS}^{-1} \Sigma_{SS^c} \beta_{S^c}\|_\infty \leq \sigma \tilde{\nu}_1 \mu_n \quad \text{and} \quad \|\Sigma_{S^c|S} \beta_{S^c}\|_\infty \leq \sigma \nu_1 \mu_n. \quad (21)$$

Notice that condition (21) is not required when β is exactly sparse, that is when it has k non-zero entries, since in this case β_{S^c} is identically equal to zero. In this case, assumptions (19, 20) for exactly sparse vectors are identical to the sufficient conditions for support recovery for the Lasso by Wainwright [26].

As an example, for the standard gaussian design, condition (19) is satisfied with $s_{\min} = s_{\max} = 1$. Condition (20) is satisfied with $\omega = 0$. Condition (21) reduces to requiring that $\max_{j \in S^c} |\beta_j| \leq \sigma \nu_1 \mu_n$, which is satisfied with $\nu_1 = \nu$.

For the case $k = 0$, instead of (19) - (21), we only make the assumption,

$$\|\Sigma \beta\|_\infty \leq \sigma \nu_1 \mu_n. \quad (22)$$

Notice that since in this case $S = \emptyset$ and $J = S^c$, alternatively, one may express the left side of the above as $\|\Sigma_{S^c|S} \beta_{S^c}\|_\infty$.

It is well known, see for example Cai and Wang [5], Tropp [23], that if the correlations between any two distinct columns are small, as given by the incoherence condition, it implies both the sparse eigenvalue condition (19) as well as the irrepresentable condition (20). We use these results to give simple sufficient conditions for (19) - (21), as well as (22) when $k = 0$, in the following lemma. For this, define the coherence parameter,

$$\gamma \equiv \gamma(\Sigma) = \max_{1 \leq j \neq j' \leq p} |\Sigma_{jj'}|. \quad (23)$$

Further, recall that $\bar{\eta} = \eta/\bar{k}$. Then we have the following.

Lemma 2.3. *Let S , with $|S| = k$, be as in (9). Assume that the correlation matrix Σ satisfies,*

$$\gamma(\Sigma) \leq \omega_0/(2\bar{k}), \quad \text{where} \quad 0 \leq \omega_0 < 1. \quad (24)$$

Further, assume that the coefficient vector β satisfies, for some $\eta \geq 0$,

$$\|\beta_{S^c}\|_1 \leq \sigma \eta \mu_n. \quad (25)$$

Define:

$$s_{\min} = 1 - \omega_0/2 \quad s_{\max} = 1 + \omega_0/2 \quad \omega = \omega_0 \quad (26)$$

$$\tilde{\nu}_1 = \omega_0 \bar{\eta} \quad \nu_1 = \nu + \omega_0 \bar{\eta}, \quad (27)$$

Then, conditions (19) - (21) holds, for $k = 1, \dots, \bar{k}$, with the above values of s_{min} , s_{max} , ω , ν_1 and $\tilde{\nu}_1$.

If $k = 0$, condition (22) holds with ν_1 in (27).

The above lemma is proved in Appendix C. Equation (24) controls the maximum correlation between distinct columns and can be regarded as the population analog of the incoherence condition (4). Condition (25) imposes that β_{S^c} has ℓ_1 norm that is $O(\eta\mu_n)$, where as mentioned before, η is allowed to scale at most linearly with \bar{k} .

Henceforth, for convenience sake, assume that we have control over the incoherence parameter as in (24) and that β satisfies (25). Further, the quantities s_{min} , s_{max} , ω , ν_1 and $\tilde{\nu}_1$ will be as in (26) and (27).

Condition (25) is more appropriate than an ℓ_1 constraint on the whole vector β since it does not impose any constraint on the larger coefficient values. Since the β_j , for $j \in S^c$, has magnitude at most $\sigma\nu\mu_n$, which is of the same order as the noise level, it makes sense for any algorithm to only estimate S accurately. In Theorem 2.4 below, we give sufficient conditions on n so that one can reliably estimate S . We note that this goal is different from that required in Zhang and Huang [27] for support recovery with approximately sparse β . There, the only constraint on β was that $\|\beta_{A_0}\|_1 = O(\eta\mu_n)$, for some set A_0 , with $|A_0^c| = k$, and where η is also allowed to grow at most linearly k . Since there was no constraint on the magnitude of β_j , for $j \in A_0$, some these β_j 's may have magnitude as high as $O(k\mu_n)$. For this reason, it made no longer sense to estimate A_0^c accurately. Their criterion for an estimate \hat{S} to be good was that $|\hat{S}| = O(k)$ and that the least squares fit of Y on the columns in \hat{S} produced a good approximation to $X\beta$.

The quantities λ_{min} , λ_{max} and λ are redefined here. These will now be expressed as functions ν , ω_0 and η using the various quantities s_{min} , s_{max} , ω , $\tilde{\nu}_1$ and ν_1 defined in (26) and (27).

We will need that the quantity $h = \sqrt{k/n} + \mu_n$ to be strictly less than one. Below, we arrange $n > 2\bar{k} \log p$. Correspondingly, one sees that $h < 1$ if, for example, $\bar{k} \geq 5$ and $p \geq 8$. Let $h_\ell = (1 - h)^2$ and $h_u = (1 + h)^2$. We define the values of λ_{min} , λ_{max} and λ in the following manner:

$$\lambda_{min} = s_{min}h_\ell \quad \text{and} \quad \lambda_{max} = s_{max}h_u. \quad (28)$$

Further,

$$\lambda = (1 + s_{max}^2\tilde{\nu}_1^2 + \nu_1\bar{\eta}) \left(1 + \bar{k}^{-1/2}\right)^2. \quad (29)$$

Let r_1 be as in (13), now replaced with the above values of λ_{min} , λ_{max} , λ . The quantity r_2 is now given by,

$$r_2 = \left[(1 - \omega) \left(\tilde{\nu}_1 + \sqrt{\frac{1 + \nu_1\bar{\eta}}{\lambda_{min}}} \right) + \sqrt{r_1} \right]. \quad (30)$$

Notice that for the i.i.d Gaussian ensemble and when β is k -sparse, the quantities ω , $\tilde{\nu}_1$, ν_1 and $\bar{\eta}$ can be taken as zero. Correspondingly, r_2 has the same form as that in (13).

Further, let $\xi = \xi(\alpha, \delta)$ be as in (12), with r_1 and r_2 appearing in its definition replaced with the values of these quantities defined above. The quantity $\tilde{p}_{err,k}$, for $k \geq 1$, which controls the probability of failure of the algorithm, is defined as,

$$\tilde{p}_{err,k} = 4/p + \frac{\sqrt{2/\pi}}{\tau} [(k+1)/p^a + k/p^{1+a}]. \quad (31)$$

We define $\tilde{p}_{err,0} = 1/p + \sqrt{(2/\pi)}/(\tau p^a)$. The threshold will now be denoted as τ_1 . It will be greater than τ by a factor $\rho \geq 1$. This factor is strictly greater than one if β is not ℓ_0 -sparse or if $\gamma(\Sigma)$ is non-zero. We are now in a position to state our main theorem.

Theorem 2.4. *Let the assumptions of Lemma 2.3 hold. Set the threshold as $\tau_1 = \rho\tau$, where τ as in (11), and*

$$\rho = \frac{\nu_1 (1 + \bar{k}^{-1/2}) + 1}{1 - \omega}. \quad (32)$$

Further, let

$$n = \xi \bar{k} \tau_1^2. \quad (33)$$

Then, if $k \geq 1$ the following holds with probability at least $1 - \tilde{p}_{err,k}$:

$$\hat{S} \subseteq S \quad \text{and} \quad \sum_{j \in \hat{F}} \beta_j^2 \leq \alpha |\hat{F}|, \quad (34)$$

where $\hat{F} = S - \hat{S}$. In particular, if $\beta_j^2 > \alpha$, for all $j \in S$, then $\hat{S} = S$ with probability at least $1 - \tilde{p}_{err,k}$.

If $k = 0$, one has that $\hat{S} = \emptyset$ with probability at least $1 - \tilde{p}_{err,0}$.

Before stating the analog of Corollary 2.2, as an aside, we give implications of the above theorem for exact recovery of support for k -sparse vectors and i.i.d designs for large n , p and k . This will help in understanding the results of Theorem 2.4 better.

In [26] it was shown that for k -sparse vectors and i.i.d Gaussian designs that there is a sharp threshold, namely $n \asymp 2k \log p$, for exact recovery of the support as n , p , k , as well as $k\beta_{min}^2/\sigma^2$, tends to infinity. This was also proved for the OMP in [14], under an additional condition on rate of increase of the signal-to-noise ratio ($\|\beta\|^2/\sigma^2$). We can get similar results using our method by recalling that for i.i.d Gaussian designs and exact sparse vectors, $s_{min} = s_{max} = 1$ and ω , ν_1 , $\tilde{\nu}_1$ and η are all zero. Further, take $\bar{k} = k$. Correspondingly, since h goes to 0, the quantities λ_{min} , λ_{max} and λ in (28, 29) tend to 1 as n , p and k become large. This implies that r_1 tends to one and r_2 (30) tends to 2. Further, as $k\beta_{min}^2/\sigma^2$ tends to infinity, one may also allow $k\alpha/\sigma^2$ tend to infinity, while keeping $\alpha < \beta_{min}$. From Theorem 2.4, this will ensure that the support will be recovered exactly. Next, let's evaluate the quantity ξ (12) appearing in the expression for n . As $k\alpha/\sigma^2$ tends to infinity, one sees that the first term in the maximum in (12) is the active one and hence ξ tends to $(1 + \delta)$ (using r_1 tends to 1). One may also appropriately choose δ to tend to zero, making ξ tend

to 1. Accordingly, from (33), one sees that if $n \approx 2(1+a)k \log p$, for large k, p , one can recover the support exactly, with probability at least $1 - \tilde{p}_{err,k}$. When β is extremely sparse, for example, when $k = O(\log p)$, then it is possible to arrange for a to decrease to 0, while making $\tilde{p}_{err,k}$ also to 0. In this case, one gets the threshold $n \approx 2k \log p$ for exact recovery. However, in the regime where k is not negligible compared to p (for example, when k/p is constant), then our results only allow for a to tend to 1 (from above), so as to ensure $\tilde{p}_{err,k}$ goes to zero. In this case our results are slightly inferior, requiring $n \approx 4k \log p$ for exact recovery. We remark in Section 6 on how the results in [14] may be carried over to the general case analyzed here.

We now state the analog of Corollary 2.2. The goal now is not to recover the non-zero entries, but only those that are large compared to the noise level, which is a subset of S . We have the following.

Corollary 2.5. *Let the assumptions of Lemma 2.3 hold and set the threshold to be τ_1 as in Theorem 2.4. Define $\bar{\xi} = 32(r_2\rho)^2(1+a)$ and $r = 2r_2\rho\sqrt{1+a}$, where r_2 as in (30). Let*

$$n \geq \bar{\xi} \bar{k} \log p. \quad (35)$$

Then, if $k \geq 1$, with probability at least $1 - \tilde{p}_{err,k}$, the estimate \hat{S} is contained in S and,

$$\left\{ j : |\beta_j| > r \sigma \sqrt{k} \mu_n \right\} \subseteq \hat{S}. \quad (36)$$

Further, if $|\beta_j| > r \sigma \mu_n$, for all $j \in S$, one has $\hat{S} = S$ with probability at least $1 - \tilde{p}_{err,k}$.

If $k = 0$, then \hat{S} is \emptyset with probability at least $1 - \tilde{p}_{err,0}$.

Corollary 2.5 gives strong performance guarantees for the OMP under an incoherence property on the correlation matrix and an ℓ_1 constraint on the smaller coefficients. From (36), one sees that the larger coefficients, that is, those with magnitude $\Omega(\sqrt{k}\mu_n)$, are contained in \hat{S} with high probability. Better performance can be demonstrated when all β_j 's, for $j \in S$, have magnitude $\Omega(\mu_n)$. In this case, it is possible to recover S , while ensuring that there are no false positives. This is in a sense ideal, since it is nearly what one would expect in the orthogonal design case discussed in the beginning of Section 2. In this case, assuming \hat{S} is as in (8), one sees that in order to prevent false positives, t needs to be $\Omega(\mu_n)$. Thus $|\beta_j|$, for $j \in S$, also needs to be $\Omega(\mu_n)$, with a slightly larger constant, to ensure $\hat{S} = S$. For example, if the $|\beta_j|$'s, for $j \in S$, is at least $\tilde{t} = (\nu + 2\sqrt{1+a})\sigma\mu_n$, then it is not hard to see that the probability $\hat{S} = S$ is at least $1 - 2/p^a$. Of course, the factor of $r\sigma$ obtained here, is larger than the corresponding factor for the orthogonal case, since the X matrix is in general quite far from being orthogonal; indeed, it is singular when $p > n$.

As a consequence of the above, we state results demonstrating strong oracle inequalities for parameter estimation under the ℓ_2 -loss.

2.2.1 Oracle inequalities under ℓ_2 -loss

Let $\hat{\beta}$ be the coefficient estimate obtained after running the algorithm. More explicitly, $(\hat{\beta}_j : j \in \hat{S})$ is simply the least squares estimate when Y is regressed on $X_{\hat{S}}$ and $\hat{\beta}_j = 0$ for $j \in \hat{S}^c$.

We assume that the correlation matrix Σ satisfies (24), that is,

$$\gamma(\Sigma) \leq \omega_0/(2\bar{k}), \quad (37)$$

where $0 \leq \omega_0 < 1$.

For simplicity, we consider the case that β satisfies (9) with $\nu = 1$, that is,

$$S = \{j : |\beta_j| > \sigma\mu_n\} \quad \text{and} \quad \|\beta_{S^c}\|_1 \leq \sigma\eta\mu_n, \quad (38)$$

where $|S| = k$ and η is allowed to grow at most linearly with \bar{k} , that is $\bar{\eta} = \eta/\bar{k}$ is $O(1)$. With $\nu = 1$, S denotes the set of indices greater than the noise level.

For the above values of η , ω_0 and with $\nu = 1$, evaluate the quantities s_{min} , s_{max} as well as $\tilde{\nu}_1$, ν_1 and ω using expressions (26) and (27). Evaluate r_2 as in (30), where the quantities λ , λ_{min} , λ_{max} are calculated using equations (28, 29). Further, let $\bar{\xi}$ and r be as in Corollary 2.5. Then we have the following.

Theorem 2.6. *Let (37) and (38) hold. For fixed such β , if*

$$n \geq \bar{\xi} \bar{k} \log p,$$

then the following holds with probability at least $1 - \tilde{p}_{err, k}$:

$$\|\hat{\beta} - \beta\|^2 \leq C \sum_{j=1}^p \min(\beta_j^2, \sigma^2 \mu_n^2), \quad (39)$$

where $C = (4/9)r^2$.

The above theorem is essentially the analog of similar results for the Lasso [29, Corollary 6.1] and Dantzig selector [6, Theorem 1.2]. Note, the latter assumes that β is k -sparse. Our results are more general since we only assume that the ℓ_1 norm of the smaller coefficients satisfies a certain bound. We proceed to state the corollary of the result assuming β is k -sparse.

For k -sparse β , we only assume that (37) holds. Take $\eta = \bar{k}$, so that $\bar{\eta} = 1$. Evaluate r_2 using this values of η , and with $\nu = 1$, and call it r_2^* , that is,

$$r_2^* = \left[(1 - \omega_0) \left(\omega_0 + \sqrt{\frac{2 + \omega_0}{\lambda_{min}}} \right) + \sqrt{r_1} \right], \quad (40)$$

where once again, the quantities r_1 and λ_{\min} as calculated using (13, 28) and equations (26) and (27). Further, let ξ^* have the same expression as $\bar{\xi}$, except it is evaluated using r_2^* instead of r_2 . Similarly, let $r^* = 2r_2^*\rho\sqrt{1+a}$. Then we have the following.

Corollary 2.7. *Let (37) hold and let β be a fixed k -sparse vector, for some $k \geq 0$. If*

$$n \geq \xi^* \bar{k} \log p,$$

then for $C_1 = (4/9)(r^)^2$, the following holds except on a set with probability $\tilde{p}_{err, k}$:*

$$\|\hat{\beta} - \beta\|^2 \leq C_1 \sum_{j=1}^p \min(\beta_j^2, \sigma^2 \mu_n^2). \quad (41)$$

We now proceed to give proofs of our main results. The proofs employs techniques developed in Zhang [28] and Tropp and Gilbert [25].

3 Proof of results in Subsection 2.1

Proof of Theorem 2.1. The following statistics will be useful in our analysis. Denote,

$$\mathcal{Z}_i = \max_{j \in S_0} |\mathcal{Z}_{ij}| \quad \text{and} \quad \tilde{\mathcal{Z}}_i = \max_{j \in S_0^c} |\mathcal{Z}_{ij}| \quad (42)$$

Notice if $\mathcal{Z}_i > \tau$ and $\tilde{\mathcal{Z}}_i > \tilde{\tau}$, then the index detected in step i , that is $a(i)$, belongs to S .

We first prove for the case $k \geq 1$. Let \mathcal{E} be the event that statement (17) in Theorem 2.1 does not hold. We want to show that the probability of \mathcal{E} is small. There are two types of errors that we wish to control. Let \mathcal{E}_1 be the event that \hat{S} is not contained in S_0 . Further, let \mathcal{E}_2 be the event that \hat{S} is contained in S_0 , however $\sum_{j \in \hat{F}} \beta_j^2 > \alpha |\hat{F}|$. Clearly, $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$.

We use an argument similar to that used in Tropp and Gilbert [25]. We initially pretend that $X = X_{S_0}$ and that the coefficient vector β is shortened to a $k \times 1$ vector β_{S_0} with all non-zero entries. Notice that $Y = X_{S_0} \beta_{S_0} + \epsilon$. For a given threshold τ , we run the algorithm on this truncated problem. Let $m \leq k$ be the number of steps and let $\tilde{R}_1, \tilde{R}_2, \dots, \tilde{R}_m$ be the associated residuals after each step. Also, denote as \tilde{R}_0 the vector Y . Notice that $m, \tilde{R}_0, \tilde{R}_1, \dots, \tilde{R}_m$ are functions of $A = [X_{S_0} : \epsilon]$.

Let \mathcal{E}_u be the event that statement (17) does not hold for the truncated problem. More explicitly, taking $\hat{S}_1 = \hat{S}(Y, X_{S_0}, \tau)$ and $\hat{F}_1 = S_0 - \hat{S}_1$, it is the event that $\|\beta_{\hat{F}_1}\|^2 > \alpha |\hat{F}_1|$.

Denote $T_i = \max_{j \in S_0} |X_j^T \tilde{R}_{i-1} / \|\tilde{R}_{i-1}\|$ and $\tilde{T}_i = \max_{j \in S_0^c} |X_j^T \tilde{R}_{i-1} / \|\tilde{R}_{i-1}\|$, for $i = 1, \dots, m+1$. Notice that the statistics T_i, \tilde{T}_i are similar to $\mathcal{Z}_i, \tilde{\mathcal{Z}}_i$, the only difference being that the residuals involved in the

former arise from running the algorithm on the truncated problem, whereas in the latter they arise from consideration of the original problem. Further, let \mathcal{E}_f be the event

$$\mathcal{E}_f = \left\{ \tilde{T}_i > \tau, \tilde{T}_i \geq T_i \text{ for some } i \leq m+1 \right\}.$$

We now show that $\mathcal{E} \subseteq \mathcal{E}_u \cup \mathcal{E}_f$. To see this, write \mathcal{E} as a disjoint union $\mathcal{E}_1 \cup \tilde{\mathcal{E}}_2$, where $\tilde{\mathcal{E}}_2 = \mathcal{E}_2 \cap \mathcal{E}_1^c$. Let's first consider the case that $\tilde{\mathcal{E}}_2$ occurs. Clearly this means that \mathcal{E}_u has occurred if the algorithm were run on the truncated problem for the given A .

Next, consider the case that \mathcal{E}_1 occurs. Let $R_0, R_1 \dots$ etc. be the residuals for the original problem (1), for the given realization of $[X : \epsilon]$. Let i^* be the step for which the false alarm occurs for the first time. Clearly, $i^* \leq m+1$, since otherwise it would mean that the truncated problem (with $X = X_{S_0}$) ran for more than m steps. Also, we must have $\{Z_i > \tau, Z_i > \tilde{Z}_i\}$ occur for $1 \leq i \leq i^* - 1$ and $\{\tilde{Z}_{i^*} > \tau, \tilde{Z}_{i^*} \geq Z_{i^*}\}$ occur. Correspondingly, one sees that $R_0 = \tilde{R}_0, \dots, R_{i^*-1} = \tilde{R}_{i^*-1}$, which implies that $T_{i^*} = Z_{i^*}$ and $\tilde{T}_{i^*} = \tilde{Z}_{i^*}$. Consequently, as $\{\tilde{T}_{i^*} > \tau, \tilde{T}_{i^*} \geq T_{i^*}\}$ occurs, \mathcal{E}_f occurs. Hence, $\mathcal{E} \subseteq \mathcal{E}_u \cup \mathcal{E}_f$ which gives,

$$\mathbb{P}(\mathcal{E}) \leq \mathbb{P}(\mathcal{E}_u) + \mathbb{P}(\mathcal{E}_f).$$

Consequently, all we are left with is to bound the probabilities of \mathcal{E}_f and \mathcal{E}_u .

We first bound the probability of \mathcal{E}_f . For this, notice that $\mathcal{E}_f \subseteq \mathcal{E}'_f$, where $\mathcal{E}'_f = \{\max_{1 \leq i \leq m+1} \tilde{T}_i > \tau\}$. Since $X_{S_0^c}$ is independent of $A = [X_{S_0} : \epsilon]$, one has that $X_{S_0^c}$ is independent of $\tilde{R}_1, \dots, \tilde{R}_m$. Correspondingly, from Lemma A.1 (a), conditional on A , we have that $X_j^T \tilde{R}_i / \|\tilde{R}_i\|$ is sub-gaussian with mean 0 and scale 1, for $j \in S_0^c$ and $1 \leq i \leq m+1$. Consequently, using standard results on the maximum of sub-Gaussian random variables (Lemma A.1 (b)), if τ be as in (11), one gets that $\mathbb{P}(\mathcal{E}_f | A) \leq 2(m+1)/p^a$, using $|S_0^c| \leq p$. Since $m \leq k$, this probability is bounded by $2(k+1)/p^a$, which implies $\mathbb{P}(\mathcal{E}_f) \leq 2(k+1)/p^a$.

Next, we bound the probability of \mathcal{E}_u . For this, consider a linear model of the form,

$$U = H\varphi + w, \tag{43}$$

where H is an $n \times k$ matrix satisfying, w an $n \times 1$ vector and φ a $k \times 1$ dimensional coefficient vector. After running the OMP on this model (with $Y = U$, $X = H$ and threshold τ_0), let $\hat{S}_2 = \hat{S}(U, H, \tau_0)$ be the estimate of the support. Further, let $\hat{\varphi}$ be the coefficient estimate obtained, that is, $(\hat{\varphi}_j : j \in \hat{S}_2)$ is the least squares estimate when U is regressed on $H_{\hat{S}_2}$ and $\hat{\varphi}_j = 0$ for j not in \hat{S}_2 . We use the following Lemma, the proof of which is similar to the analysis in Zhang [28].

Lemma 3.1. *For the model (43), let the following hold.*

- (i) *Condition 1 holds for H , that is the eigenvalues of $H^T H/n$ are between λ_{\min} and λ_{\max} .*
- (ii) *Condition 2 holds for w , that is $\|w\|^2 \leq n\sigma^2\lambda$, for some $\lambda > 0$.*

(iii) $\|\hat{\varphi}_{ls} - \varphi\|_\infty \leq \sigma c_0 \tau_0 / \sqrt{n}$, for some constant $c_0 > 0$, where $\hat{\varphi}_{ls}$ is the coefficient vector of the least square fit of U on H .

Under the above, if the OMP is run with $Y = U$, $X = H$ and threshold τ_0 , when the algorithm stops we must have the following,

(a)

$$\left(1 - \tau_0 \sqrt{r_1 k/n}\right) \|\varphi_{\hat{F}_2}\| \leq \tilde{r}_2 \sigma \tau_0 \sqrt{\frac{|\hat{F}_2|}{n}}, \quad (44)$$

where $\hat{F}_2 = \{1, \dots, k\} - \hat{S}_2$, denotes the indices not detected after running the algorithm. Further, r_1 has the same form as (13), replaced with the above values of λ_{min} , λ_{max} and λ . Also, $\tilde{r}_2 = c_0 + \sqrt{r_1}$.

(b)

$$\|\hat{\varphi} - \varphi\| \leq \frac{\tilde{r}_2 \sigma \tau_0 \sqrt{k/n}}{1 - \tau_0 \sqrt{r_1 k/n}}. \quad (45)$$

The above lemma is proved in Appendix B. We only require the conclusions in part (a) of the lemma for the time being. Part (b) will be required of Subsection 2.2.1 to get bounds on ℓ_2 -error of the coefficient estimate.

Now apply Lemma 3.1 to the truncated problem, that is, with $H = X_{S_0}$, $\varphi = \beta_{S_0}$, $U = Y$ and $\tau_0 = \tau$. Notice that in this case $\hat{F}_2 = \hat{F}_1$ and $\hat{S}_2 = \hat{S}_1$. We know that requirements (i) and (ii) of the Lemma 3.1 hold, except on a set \mathcal{E}_{cond} . The following lemma shows that (iii) holds with high probability.

Lemma 3.2. *Let $\hat{\beta}_{ls}$ be the least squares fit when Y is regressed on X_{S_0} . Further, let*

$$\mathcal{E}_{ls} = \{\|\hat{\beta}_{ls} - \beta_{S_0}\|_\infty > \sigma c_0 \tau / \sqrt{n}\},$$

where $c_0 = 1/\sqrt{\lambda_{min}}$. Then $\mathbb{P}(\mathcal{E}_{ls} \cap \mathcal{E}_{cond}) \leq 2k/p^{1+a}$.

The above lemma is proved after this proof. Using the above lemma, all requirements of Lemma 3.1 hold, except on a set $\tilde{\mathcal{E}}_u = \mathcal{E}_{cond} \cup \mathcal{E}_{ls}$, the probability of which is bounded by $\mathbb{P}(\mathcal{E}_{cond}) + 2k/p^{1+a}$. We now show that $\mathcal{E}_u \subseteq \tilde{\mathcal{E}}_u$. We do this by showing $\tilde{\mathcal{E}}_u^c \subseteq \mathcal{E}_u^c$. To see this, notice that on $\tilde{\mathcal{E}}_u^c$, one has

$$\left(1 - \tau \sqrt{r_1 k/n}\right) \|\beta_{\hat{F}_1}\| \leq \tilde{r}_2 \sigma \tau \sqrt{\frac{|\hat{F}_1|}{n}}. \quad (46)$$

from (44). Assume that \hat{F}_1 is non-empty, since otherwise the claim is trivially true. Notice that since $n \geq (1 + \delta)r_1 \bar{k} \tau^2$ from (16), one has $\tau (\bar{k} r_1 / n)^{1/2} \leq 1/\sqrt{1 + \delta}$. Now, since $k \leq \bar{k}$, the left side of (46) is non-negative. Thus, (46) can be reexpressed as,

$$\|\beta_{\hat{F}_1}\|^2 \leq (\sigma^2 r_2^2 f(\delta) \tau^2 / n) |\hat{F}_1|,$$

which follows from noticing that $r_2 = \tilde{r}_2$, where r_2 is as in (13). Now, since $n \geq \sigma^2 r_2^2 f(\delta) \tau^2 / \alpha$, the left side of the above is at most $\alpha |\hat{F}_1|$. Thus, $\sum_{j \in \hat{F}_1} \beta_j^2 \leq \alpha |\hat{F}_1|$ on $\tilde{\mathcal{E}}_u^c$, which implies that $\mathcal{E}_u \subseteq \tilde{\mathcal{E}}_u$. Consequently, $\mathbb{P}(\mathcal{E}_u) \leq \mathbb{P}(\mathcal{E}_{cond}) + 2k/p^{1+a}$. Accordingly, since $\mathbb{P}(\mathcal{E}) \leq \mathbb{P}(\mathcal{E}_u) + \mathbb{P}(\mathcal{E}_f)$, one has $\mathbb{P}(\mathcal{E}) \leq \mathbb{P}(\mathcal{E}_{cond}) + 2k/p^{1+a} + 2(k+1)/p^a$, which is equal to $p_{err,k}$. This completes the proof for the case $k \geq 1$.

For the case $k = 0$, we just need to show that the algorithm stops after the first step, in which case $\hat{S} = \emptyset$. This is immediately seen by noticing that for $k = 0$, one has that \mathcal{Z}_{1j} , for $j \in J$, are sub-gaussian with mean 0 and scale 1. Correspondingly, from Lemma A.1(b), the event $\{\max_{j \in J} |\mathcal{Z}_{1j}| > \tau\}$ has probability at most $p_{err,0} = 2/p^a$. \square

Proof of Lemma 3.2. Firstly, note that $\hat{\beta}_{ls} - \beta_{S_0}$ can be expressed as $Z = (X_{S_0}^\top X_{S_0})^{-1} X_{S_0}^\top \epsilon$. Let $Z = (Z_j : j = 1, \dots, k)$. Now, conditioned on X_{S_0} , each Z_j is sub-gaussian with mean 0 and scale $\sigma_j = \sigma \sqrt{e_j^\top (X_{S_0}^\top X_{S_0})^{-1} e_j}$. Here, e_j is the j th column of the size k identity matrix. Correspondingly, from Lemma A.1(b), one gets $\max_j |Z_j|$ is less than $(\max_j \sigma_j) \tau$, except on a set with probability $2k/p^{1+a}$. Finally, observe that on \mathcal{E}_{cond}^c , one has $e_j^\top (X_{S_0}^\top X_{S_0})^{-1} e_j \leq 1/(n \lambda_{min})$, since the maximum eigenvalue of $(X_{S_0}^\top X_{S_0}/n)^{-1}$ is at most $1/\lambda_{min}$. Thus, $\max_j \sigma_j \tau$ is at most $\sigma c_0 \tau / \sqrt{n}$, with $c_0 = 1/\sqrt{\lambda_{min}}$. \square

Proof of Corollary 2.2. Take $\alpha(\delta) = \sigma^2 / [(1 + \delta) \bar{k}]$. Further, let $\xi(\delta) = \xi(\alpha(\delta), \delta)$, which, using $r_2^2 \geq r_1$ and $f(\delta) \geq 1$, can be written as,

$$\xi(\delta) = (1 + \delta) f(\delta) r_2^2. \quad (47)$$

The function $(1 + \delta) f(\delta)$, for $\delta > 0$, has its minimum at $\delta^* = 3$. Further, it is increasing and goes to infinity as δ tends to infinity. Now, using $\xi(\delta^*) = 16r_2^2$, notice that $\xi(\delta^*) \bar{k} \tau^2 = \bar{\xi} \bar{k} \log p$. Correspondingly, since $n \geq \bar{\xi} \bar{k} \log p$, one gets that

$$n = \xi(\delta) \bar{k} \tau^2, \quad (48)$$

for some $\delta \geq \delta^*$. Consequently, from Theorem 2.1, one has,

$$\hat{S} \subseteq S_0 \quad \text{and} \quad \sum_{j \in \hat{F}} \beta_j^2 \leq \alpha(\delta) |\hat{F}|, \quad (49)$$

with probability at least $1 - p_{err,k}$. Use $f(\delta) \leq f(\delta^*) = 4$, to get from (48) that $n \leq (1 + \delta) r \bar{k} \tau^2$. Correspondingly, $\alpha(\delta)$ is at most $r^2 \sigma^2 \mu_n^2$. Consequently, any j , with $|\beta_j| > r \sigma \sqrt{\bar{k} \mu_n}$ cannot be in \hat{F} since it would contradict the inequality in (49). Further, if $\beta_{min} > r \sigma \mu_n$, the inequality in (49) cannot hold if \hat{F} is non-empty. In this case the algorithm recovers the entire support. \square

4 Proof of results in Subsection 2.2

Proof of Theorem 2.4. Once again, we first prove for the case $k \geq 1$. As before, we are interested in bounding the probability of \mathcal{E} , where $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$. Here \mathcal{E}_1 is the event that \hat{S} is not contained in $S = S(\beta)$. Also, \mathcal{E}_2 is the event $\hat{S} \subseteq S$ and $\|\beta_{\hat{F}}\|^2 > \alpha|\hat{F}|$, where, here $\hat{F} = S - \hat{S}$ and $\hat{S} = \hat{S}(Y, X, \tau_1)$. Write Y as $Y = X_S \beta_S + \tilde{\epsilon}$, where $\tilde{\epsilon} = X_{S^c} \beta_{S^c} + \epsilon$. Analogous to before, we initially pretend that $X = X_S$ and $\beta = \beta_S$ and run the algorithm on the truncated problem to get residuals $\tilde{R}_0, \tilde{R}_1, \tilde{R}_2, \dots, \tilde{R}_m$. These residuals are functions of $A = [X_S : \tilde{\epsilon}]$. Further, as before, let \mathcal{E}_u be the event that statement (34) is not met for this truncated problem. With $\hat{S}_1 = \hat{S}(Y, X_S, \tau_1)$ and $\hat{F}_1 = S - \hat{S}_1$, it is the event that $\|\beta_{\hat{F}_1}\|^2 > \alpha|\hat{F}_1|$. Similarly, we define T_i, \tilde{T}_i as before, now with the maximum taken over S instead of S_0 . Further, define the event \mathcal{E}_f analogous to before, with τ replaced by τ_1 . Using the same reasoning as in Theorem 2.1, one has $\mathcal{E} \subseteq \mathcal{E}_u \cup \mathcal{E}_f$. We first proceed to bound the probability of \mathcal{E}_f . Notice that unlike previously, the X_j 's, for $j \in S^c$, are not independent of the \tilde{R}_i 's. This makes bounding the probability of \mathcal{E}_f more involved.

The following lemma will be useful, both in bounding $\mathbb{P}(\mathcal{E}_f)$ as well as $\mathbb{P}(\mathcal{E}_u)$. We denote as $\hat{\beta}_{ls}$ the least square estimate when Y is regressed on X_S .

Lemma 4.1. *Parts (i)-(iii) of this lemma demonstrate that requirements (i)-(iii) of Lemma 3.1 are satisfied with high probability.*

(i) *With $\lambda_{min}, \lambda_{max}$ as in (28), the following holds with probability at least $1 - 2/p$:*

$$\lambda_{min}\|v\|^2 \leq \|X_S v\|^2/n \leq \lambda_{max}\|v\|^2 \quad \text{for all } v \in \mathbb{R}^k. \quad (50)$$

(ii) *Let λ be as in (29). Then $\|\tilde{\epsilon}\|^2/(n\sigma^2) \leq \lambda$, with probability at least $1 - 1/p$.*

(iii) *Let $\mathcal{E}_{ls} = \{\|\hat{\beta}_{ls} - \beta_S\|_\infty > \sigma c_0 \tau_1 / \sqrt{n}\}$, where*

$$c_0 = (1 - \omega) \left[\tilde{\nu}_1 + \sqrt{\frac{1 + \nu_1 \bar{\eta}}{\lambda_{min}}} \right] \quad (51)$$

Then $\mathbb{P}(\mathcal{E}_{cond}^c \cap \mathcal{E}_{ls}) \leq (\sqrt{2/\pi})k/(\tau p^{1+a})$, where \mathcal{E}_{cond} , here, is the event that (i) or (ii) above fails. From (i) and (ii) it has probability at most $3/p$.

The above lemma is proved in Section 5. As mentioned before, the X_j 's, for $j \in S^c$, are not independent of the \tilde{R}_i 's. We get around this by finding the conditional distribution of each X_j given X_S and $\tilde{\epsilon}$. Correspondingly, each X_j may be represented as a linear combination of columns in $A = [X_S : \tilde{\epsilon}]$ plus a noise vector, which we call Z_j . This noise term is independent of A and hence $\tilde{R}_0, \tilde{R}_1, \dots, \tilde{R}_m$.

Let $a_j = \Sigma_{SS}^{-1} \Sigma_{Sj}$ and

$$b_j = \frac{e_j^T \Sigma_{S^c|S} \beta_{S^c}}{\sqrt{d}}, \quad (52)$$

where e_j is the j th column of the size $p - k$ identity matrix and

$$d = \sigma^2 + \beta_{S^c}^T \Sigma_{S^c|S} \beta_{S^c}. \quad (53)$$

The following lemma characterizes the conditional distribution of X_j given A .

Lemma 4.2. *Let a_j, b_j , for $j \in S^c$, be as above. Then we have the following:*

(i) *The distribution of X_j , for $j \in S^c$, may be represented as*

$$X_j \stackrel{\mathcal{D}}{=} X_S a_j + b_j W + Z_j \quad (54)$$

where $W \sim N(0, I_n)$ and is independent of X_S . Further, Z_j is independent of $[X_S : \tilde{\epsilon}]$ and follows $N(0, \tilde{\sigma}_{jj} I_n)$, with $\tilde{\sigma}_{jj} \leq \sigma_{jj} = 1$.

(ii) *Define, for $j \in S^c$ and $i = 1, \dots, m + 1$,*

$$V_{ji} = b_j W^T \frac{\tilde{R}_{i-1}}{\|\tilde{R}_{i-1}\|} + E_{ji}, \quad (55)$$

where $E_{ji} = Z_j^T \tilde{R}_{i-1} / \|\tilde{R}_{i-1}\|$. Let,

$$\tilde{\mathcal{E}}_f = \left\{ \max_{1 \leq i \leq m+1, j \in S^c} |V_{ji}| > (1 - \omega)\tau_1 \right\}. \quad (56)$$

Then $\mathbb{P}(\tilde{\mathcal{E}}_f) \leq 1/p + (\sqrt{2/\pi})(k + 1)/(\tau p^a)$.

The above lemma is proved in Section 5. We now show that $\mathcal{E}_f \subseteq \tilde{\mathcal{E}}_f$. To see this, notice that on $\tilde{\mathcal{E}}_f^c$ one has,

$$\begin{aligned} \tilde{T}_i &\leq (\max_{j \in S^c} \|a_j\|_1) T_i + (1 - \omega)\tau_1 \\ &\leq \omega T_i + (1 - \omega)\tau_1, \end{aligned} \quad (57)$$

for $i = 1, \dots, m + 1$. Here, the first inequality follows from using (54) and $\left| a_j^T X_S^T \tilde{R}_{i-1} / \|\tilde{R}_{i-1}\| \right| \leq \|a_j\|_1 T_i$, along with the fact that $|V_{ji}|$ is bounded by $(1 - \omega)\tau_1$ on $\tilde{\mathcal{E}}_f^c$. The second inequality follows from (20). We now show that

$$\mathcal{E}' = \left\{ \tilde{T}_i \leq \omega T_i + (1 - \omega)\tau_1 \text{ for each } i \leq m + 1 \right\}$$

implies \mathcal{E}_f^c . To see this, for each i , consider two cases, viz. $T_i > \tau_1$ and $T_i \leq \tau_1$. From (57), in the first case one has $\tilde{T}_i < T_i$, and in the second case, one has $\tilde{T}_i \leq \tau_1$. Correspondingly, \mathcal{E}' is contained in

$$\{\tilde{T}_i < T_i \text{ or } \tilde{T}_i \leq \tau_1 \text{ for each } i \leq m + 1\},$$

which is \mathcal{E}_f^c . Consequently, $\mathcal{E}_f \subseteq \tilde{\mathcal{E}}_f$. Consequently, $\mathbb{P}(\mathcal{E}_f) \leq 1/p + (\sqrt{2/\pi})(k + 1)/(\tau p^a)$ from Lemma 4.2.

What remains to be seen is that the probability of the event \mathcal{E}_u can be bounded as before. For this we apply Lemma 3.1 once again. That conditions (i) - (iii), required for application of Lemma 3.1, are satisfied with high probability is proved parts (i)-(iii) of Lemma 4.1. Consequently, as before, if $\tilde{\mathcal{E}}_u = \mathcal{E}_{cond} \cup \mathcal{E}_{ls}$, where the sets on the right side are as in Lemma 4.1, one gets that on $\tilde{\mathcal{E}}_u^c$,

$$\left(1 - \tau_1 \sqrt{r_1 k/n}\right) \|\beta_{\hat{F}_1}\| \leq \tilde{r}_2 \sigma \tau_1 \sqrt{\frac{|\hat{F}_1|}{n}}. \quad (58)$$

Here $\tilde{r}_2 = c_0 + \sqrt{r_1}$, where c_0 as in (51). Notice that $\tilde{r}_2 = r_2$, where r_2 as in (30). Now, once again use the fact that $n \geq (1 + \delta)r_1 k \tau_1^2$ and $n \geq r_2^2 f(\delta) \sigma^2 \tau_1^2 / \alpha$, to get that (58) implies \mathcal{E}_u^c . Accordingly, $\mathbb{P}(\mathcal{E}_u) \leq \mathbb{P}(\tilde{\mathcal{E}}_u)$. Consequently, one has,

$$\begin{aligned} \mathbb{P}(\mathcal{E}) &\leq \mathbb{P}(\mathcal{E}_u \cup \mathcal{E}_f) \\ &\leq \mathbb{P}(\mathcal{E}_{cond}) + \mathbb{P}(\mathcal{E}_{cond}^c \cap \mathcal{E}_{ls}) + \mathbb{P}(\mathcal{E}_f), \end{aligned}$$

which is at most $\tilde{p}_{err,k} = 4/p + (\sqrt{2/\pi}/\tau) [(k+1)/p^a + k/p^{1+a}]$. This completes the proof for $k \geq 1$.

If $k = 0$, we will show that the probability that $\max_{j \in J} |\mathcal{Z}_{1j}|$ exceeds τ_1 is at most $\tilde{p}_{err,0}$. This would imply that the algorithm stops after one step and \hat{S} is empty. Notice that $S^c = J$ and hence $\tilde{\epsilon} = Y$. Consequently, $X_j \stackrel{\mathcal{D}}{=} \tilde{b}_j Y / \sigma_Y + Z_j$, where $Z_j \sim N(0, \tilde{\sigma}_j)$ is independent of Y , with $\tilde{\sigma}_j \leq 1$. Also, $\tilde{b}_j = e_j^T \Sigma \beta / \sigma_Y$, where $\sigma_Y^2 = \text{Var}(Y_1) = \sigma^2 + \beta^T \Sigma \beta$. Correspondingly,

$$\mathcal{Z}_{1j} \stackrel{\mathcal{D}}{=} \tilde{b}_j \|Y\| / \sigma_Y + Z_j^T \frac{Y}{\|Y\|} \quad (59)$$

Using $\sigma_Y \geq \sigma$, one has $\tilde{b}_j \leq \nu_1 \mu_n$. Further, using $\|Y\| / \sigma_Y \leq (1 + \mu_n)$, with probability at least $1 - 1/p$ from Lemma A.2, one has that the first term in the right side of (59) is at most $\nu_1 \tau (1 + \bar{k}^{-1/2})$ with probability at least $1 - 1/p$. Further $|Z_j^T Y / \|Y\||$, using the independence of Z_j and Y , is less than τ for all j with probability at least $1 - \sqrt{2/\pi}/(\tau p^a)$ (Lemma A.1 (b)). Denoting, $\tau_2 = [\nu_1 (1 + \bar{k}^{-1/2}) + 1] \tau$, one sees $\max_{j \in J} |\mathcal{Z}_{1j}| \leq \tau_2$, with probability at least $1 - \tilde{p}_{err,0}$. Notice that since $\tau_1 \geq \tau_2$, the event $\max_{j \in J} |\mathcal{Z}_{1j}| \leq \tau_1$ also has probability at least $1 - \tilde{p}_{err,0}$. This completes the proof. \square

Proof of Corollary 2.5. The proof is exactly similar to that of Corollary 2.2. As before, taking $\alpha(\delta) = \sigma^2 / [(1 + \delta)\bar{k}]$ and $\xi(\delta) = \xi(\alpha(\delta), \delta)$, we notice that $\rho^2 \xi(\delta^*) \bar{k} \tau^2 = \bar{\xi} \bar{k} \log p$, where $\delta^* = 3$. Correspondingly, if $n \geq \bar{\xi} \bar{k} \log p$, one has $n = \rho^2 \xi(\delta) \bar{k} \tau^2$ for some $\delta \geq \delta^*$ and hence,

$$\hat{S} \subseteq S \quad \text{and} \quad \sum_{j \in \hat{F}} \beta_j^2 \leq \alpha(\delta) |\hat{F}|$$

with probability at least $1 - \tilde{p}_{err,k}$, from Theorem 2.4. Further, $\alpha(\delta)$ is at most $r^2 \sigma^2 \mu_n^2$, using the same reasoning as before. The conclusions on recovering the large coefficients follow immediately from this. \square

Proof of Theorem 2.6. Notice that,

$$\|\hat{\beta} - \beta\|^2 = \|\hat{\beta}_S - \beta_S\|^2 + \|\hat{\beta}_{S^c} - \beta_{S^c}\|^2. \quad (60)$$

We apply the result of Corollary 2.5, to get that except on a set with probability $\tilde{p}_{err,k}$, one has $\hat{S} \subseteq S$. Correspondingly, the second term in (60) is simply $\|\beta_{S^c}\|^2$, which is equal to $\sum_{j \in S^c} \min\{\beta_j^2, \sigma^2 \mu_n^2\}$.

Let's next concentrate on the first term in (60). Notice that since $\hat{S} \subseteq S$, one has $\hat{\beta}_S$ is same as the coefficient estimate one would get if the OMP were run on the truncated problem. Correspondingly, using part (b) of Lemma 3.1, with $\tau_0 = \tau_1$ and $\tilde{r}_2 = r_2$, one gets that

$$\|\hat{\beta}_S - \beta_S\| \leq \frac{r_2 \sigma \tau_1 \sqrt{k/n}}{1 - \tau_1 \sqrt{r_1 k/n}}, \quad (61)$$

with probability at least $1 - \tilde{p}_{err,k}$. Next, use the fact that $\tau_1 \sqrt{k/n} \leq 1/(4r_2)$ using $\bar{\xi} \bar{k} \log p = 16r_2^2 \bar{k} \tau_1^2$. Consequently, the denominator in the right side of (61) is at least $1 - \sqrt{r_1}/4r_2$. The latter is at least $3/4$ using $r_2 \geq \sqrt{r_1}$. Thus,

$$\begin{aligned} \|\hat{\beta}_S - \beta_S\| &\leq \frac{4r_2 \rho \sqrt{1+a}}{3} \sigma \sqrt{k} \mu_n, \\ &= \sqrt{C} \sigma \sqrt{k} \mu_n, \end{aligned} \quad (62)$$

where $C = (4/9)r^2$. Correspondingly, from (60) one gets that,

$$\begin{aligned} \|\hat{\beta} - \beta\|^2 &\leq C \sigma^2 k \mu_n^2 + \sum_{j \in S^c} \min\{\beta_j^2, \sigma^2 \mu_n^2\} \\ &\leq C \sum_{j=1}^p \min\{\beta_j^2, \sigma^2 \mu_n^2\}, \end{aligned}$$

where the last inequality from using $\sigma^2 k \mu_n^2 = \sum_{j \in S} \min\{\beta_j^2, \sigma^2 \mu_n^2\}$, since $S = \{j : |\beta_j| > \sigma \mu_n\}$. \square

Proof of Corollary 2.7. For k -sparse β , once again let $S = \{j : |\beta_j| > \sigma \mu_n\}$. Now $\|\beta_{S^c}\|_1 \leq \eta \sigma \mu_n$, where $\eta = \bar{k}$, since there are at most \bar{k} non-zero entries outside of S , with magnitude at most $\sigma \mu_n$. Now apply Theorem 2.6, with $\eta = \bar{k}$ (or $\bar{\eta} = 1$) to get the desired result. \square

5 Proof of results from Section 4

The following simple lemma will prove useful in proving Lemma 4.1.

Lemma 5.1. *Let $\theta_n = \bar{k}^{1/2} \mu_n$. Conditions (19) - (21) imply the following:*

- (i) *Let d be as in (53). Then $d \leq \sigma^2(1 + \nu_1 \bar{\eta} \theta_n^2)$.*
- (ii) *$\|\Sigma_{SS} g\|^2 \leq \sigma^2 s_{max}^2 \tilde{\nu}_1^2 \theta_n^2$, where $g = \Sigma_{SS}^{-1} \Sigma_{SS^c} \beta_{S^c}$.*

Remark: Since we take $n > 2\bar{k} \log p$, we have $\theta_n \leq 1$. Accordingly, the above bound holds with θ_n replaced by 1.

Proof of Lemma 5.1. We first prove part (i). Recall that $d = \sigma^2 + \beta_{S^c}^\top \Sigma_{S^c|S} \beta_{S^c}$. Write $\beta_{S^c}^\top \Sigma_{S^c|S} \beta_{S^c}$ as $\sum_{j \in S^c} \beta_j e_j^\top \Sigma_{S^c|S} \beta_{S^c}$, which can be bounded by $(\|\Sigma_{S^c|S} \beta_{S^c}\|_\infty) \|\beta_{S^c}\|_1$, which is at most $\sigma \nu_1 \bar{\eta} \theta_n^2$ from (21) and (10). This completes the proof.

For part (ii) use the fact that $\|\Sigma_{SS} g\|^2 \leq s_{max}^2 \|g\|^2$ from (19) and $\|g\| \leq \sigma \sqrt{k} \tilde{\nu}_1 \mu_n$ from (21), to complete the proof. \square

Proof of Lemma 4.1. We use a result in Szarek [21] that gives tails bounds for the largest and smallest singular values of Gaussian random matrices. Let $U \in \mathbb{R}^{n \times k}$ be a matrix with i.i.d. standard Gaussian entries. Then, for $r > 0$, one has,

$$\mathbb{P}(\lambda_k(U/\sqrt{n}) > 1 + \sqrt{k/n} + r) \leq e^{-nr^2/2}$$

$$\mathbb{P}(\lambda_1(U/\sqrt{n}) < 1 - \sqrt{k/n} - r) \leq e^{-nr^2/2},$$

where $\lambda_k(\cdot)$ and $\lambda_1(\cdot)$ gives the largest and smallest singular values respectively, of an $n \times k$ matrix. Now, taking $r = \mu_n$, one has, using the above, that with probability at $1 - 2/p$ the following holds:

$$h_\ell \|v\|^2 \leq \frac{1}{n} \|Uv\|^2 \leq h_u \|v\|^2 \quad \text{for all } v \in \mathbb{R}^k.$$

Now, notice that since $X_S \stackrel{\mathcal{D}}{=} U \Sigma_{SS}^{1/2}$, one has from the above that, with probability at least $1 - 2/p$,

$$h_\ell \|\Sigma_{SS}^{1/2} v\|^2 \leq \frac{1}{n} \|X_S v\|^2 \leq h_u \|\Sigma_{SS}^{1/2} v\|^2 \quad \text{for all } v \in \mathbb{R}^k.$$

Correspondingly, from (19), since $s_{min} \leq \|\Sigma_{SS}^{1/2} v\|^2 / \|v\|^2 \leq s_{max}$, which implies that, with probability at least $1 - 2/p$,

$$\lambda_{min} \|v\|^2 \leq \frac{1}{n} \|X_S v\|^2 \leq \lambda_{max} \|v\|^2 \quad \text{for all } v \in \mathbb{R}^k,$$

where λ_{min} , λ_{max} as in (29).

Before proving parts (ii) and (iii), observe that by conditioning on X_S , the distribution of $\tilde{\epsilon}$ may be expressed as,

$$\tilde{\epsilon} \stackrel{\mathcal{D}}{=} X_S g + \sqrt{d} W, \tag{63}$$

where $g = \Sigma_{SS}^{-1} \Sigma_{SS^c} \beta_{S^c}$ and d as in (53). Here $W \sim N(0, I_n)$ and is independent of X_S .

For part (ii), notice that from the above $\tilde{\sigma}^2 := \text{Var}(\tilde{\epsilon}_1) = \|\Sigma_{SS} g\|^2 + d$, which is at most $\sigma^2(1 + s_{max}^2 \tilde{\nu}_1^2 + \nu_1 \bar{\eta})$ from Lemma 5.1. Further, $\|\tilde{\epsilon}\|^2 / \tilde{\sigma}^2 \sim \mathcal{X}_n^2$. Now from Lemma A.2, the probability of the event $\|\tilde{\epsilon}\|^2 / (n \tilde{\sigma}^2) >$

$(1 + \mu_n)^2$ is bounded $1/p$. Use $\mu_n \leq \bar{k}^{-1/2}$ and $\tilde{\sigma}^2 \leq \sigma^2(1 + s_{max}^2 \tilde{\nu}_1^2 + \nu_1 \bar{\eta})$, to get that $\mathbb{P}(\|\tilde{\epsilon}\|^2/(n\sigma^2) > \lambda) \leq 1/p$, where λ as in (29).

For part (iii), notice that $\hat{\beta}_{ls} - \beta_S = (X_S^T X_S)^{-1} X_S^T \tilde{\epsilon}$, which using (63), can be expressed as,

$$\hat{\beta}_{ls} - \beta_S \stackrel{\mathcal{D}}{=} g + \sqrt{d}(X_S^T X_S)^{-1} X_S^T W. \quad (64)$$

Let $\tilde{\mathcal{E}}_{ls} = \{\sqrt{d}\|(X_S^T X_S)^{-1} X_S^T W\|_\infty > \sigma\sqrt{1 + \nu_1 \bar{\eta}}/\sqrt{\lambda_{min} n}\}$. Now, since W is independent of X_S , and $d \leq \sigma^2(1 + \nu_1 \bar{\eta})$, one can use the same logic as in the proof of Lemma 3.2 to get that, $\mathbb{P}(\mathcal{E}_{cond}^c \cap \tilde{\mathcal{E}}_{ls}) \leq \sqrt{2/\pi k}/(\tau p^{1+a})$. Further, $\|g\|_\infty \leq \sigma \tilde{\nu}_1 \mu_n$ using (21), which, using $\mu_n \leq \tau/\sqrt{n}$, is at most $\sigma \tilde{\nu}_1 \tau/\sqrt{n}$. Accordingly, on $\mathcal{E}_{cond}^c \cap \tilde{\mathcal{E}}_{ls}^c$, one has,

$$\begin{aligned} \|\hat{\beta}_{ls} - \beta_S\|_\infty &\leq \sigma \left[\tilde{\nu}_1 + \sqrt{\frac{1 + \nu_1 \bar{\eta}}{\lambda_{min}}} \right] \tau/\sqrt{n}, \\ &= \sigma \frac{c_0}{\sqrt{n}} \frac{\tau}{1 - \omega}, \end{aligned}$$

where c_0 as in (51). Now use $\tau/(1 - \omega) \leq \tau_1$, to get that $\mathbb{P}(\mathcal{E}_{cond}^c \cap \mathcal{E}_{ls}) \leq \sqrt{2/\pi k}/(\tau p^{1+a})$. This completes the proof of the lemma. \square

Proof of Lemma 4.2. We first prove part (i). Recall, from (63), one has, $\tilde{\epsilon} \stackrel{\mathcal{D}}{=} X_S g + \sqrt{d}W$, where $g = (\Sigma_{SS})^{-1} \Sigma_{SS^c} \beta_{S^c}$ and d as in (53). Further, W is independent of X_S and follows $N(0, I_n)$. Correspondingly, the conditional distribution of X_j given $[X_S : W]$ may be expressed as,

$$X_j \stackrel{d}{=} X_S a_j + b_j W + Z_j$$

where $a_j = \text{Cov}(X_{1,S}, X_{1j})[\text{Var}(X_{1,S})]^{-1}$ and $b_j = \text{Cov}(X_{1j}, W_1)$. Further, $Z_j \sim N(0, \tilde{\sigma}_{jj} I_n)$ and is independent of X_S and W , with

$$\tilde{\sigma}_{jj} = \sigma_{jj} - a_j^T \Sigma_{SS} a_j - b_j^2,$$

which is at most 1. Clearly, the expression for a_j matches that given in the statement of the lemma. Further, from (63), one has that,

$$\text{Cov}(X_{1j}, W_1) = \frac{1}{\sqrt{d}} [\text{Cov}(X_{1j}, \tilde{\epsilon}_1) - \text{Cov}(X_{1j}, X_{1,S} g)].$$

Notice that $\text{Cov}(X_{1j}, \tilde{\epsilon}_1) = \Sigma_{jS^c} \beta_{S^c}$ and $\text{Cov}(X_{1j}, X_{1,S} g) = \text{Cov}(X_{1j}, X_{1,S}) g$, which is $\Sigma_{jS} \Sigma_{SS}^{-1} \Sigma_{SS^c} \beta_{S^c}$. Correspondingly, the numerator of the above is $e_j^T \Sigma_{S^c|S} \beta_{S^c}$, and hence, the expression for b_j given above matches that in (52).

We now prove part (ii) of Lemma 4.2. Firstly, notice that $\max_{j \in S^c} |b_j| \leq \nu_1 \mu_n$. This follows from observing that $d \geq \sigma^2$, from (53), and also the fact that $|e_j^T \Sigma_{S^c|S} \beta_{S^c}| \leq \sigma \nu_1 \mu_n$, for all $j \in S^c$, from (21).

Recall the statistic V_{ji} given by (55). One sees that,

$$|V_{ji}| \leq |b_j| \|W\| + |E_{ji}|. \quad (65)$$

Now $\|W\|^2 \sim \mathcal{X}_n^2$. Correspondingly, from Lemma A.2, the event $\{\|W\|/\sqrt{n} > (1 + \mu_n)\}$ has probability at most $1/p$.

Further, Z_j 's are independent of $[X_S : \tilde{\epsilon}]$ and, hence, are also independent of $\tilde{R}_0, \dots, \tilde{R}_m$, since these residuals are functions of $[X_S : \tilde{\epsilon}]$. Consequently, the E_{ji} 's are standard normal random variables; Indeed, conditional on the \tilde{R}_i 's, they follow $N(0, 1)$, and hence, follow the same distribution unconditionally. Accordingly, using the same logic as in the proof of Theorem 2.1, the event

$$\left\{ \max_{1 \leq i \leq m+1, j \in S^c} |E_{ji}| > \tau \right\} \quad (66)$$

has probability bounded by $\sqrt{2/\pi}(k+1)/(\tau p^a)$.

Consequently, using the bounds on $|b_j|$ and the above, one gets that except on a set with probability $1/p + \sqrt{2/\pi}(k+1)/(\tau p^a)$, one has

$$\max_{1 \leq i \leq m+1, j \in S^c} |V_{ji}| \leq \nu_1 \mu_n \sqrt{n} (1 + \mu_n) + \tau.$$

Using $\tau \geq \mu_n \sqrt{n}$ and $\mu_n \leq \bar{k}^{-1/2}$, the right side of the above is at most $(1 - \omega)\tau_1$. This completes the proof of the lemma. \square

6 Conclusion

The paper analyzed variable selection for the OMP for random X matrices. We analyzed performance with i.i.d sub-Gaussian designs, which has uses in compressed sensing. We remark that for these i.i.d designs, the analysis carries over for the hard thresholded version of the algorithm, in which, instead of choosing the j which maximizes the $|\mathcal{Z}_{ij}|$'s, one chooses all j satisfying $|\mathcal{Z}_{ij}| > \tau$. It is only when there is some correlation within the rows that we find it advantageous to choose the index which maximizes $|\mathcal{Z}_{ij}|$.

For Gaussian designs, with correlation within rows, we give much more general results. Apart from showing that results similar to that in [26], for exact support recovery, are also possible using the OMP, we show additional recovery properties by relaxing the assumption of exact sparsity to a more realistic assumption of a control over the ℓ_1 -norm of the smaller coefficients. Oracle inequalities for the coefficient estimate also followed easily as a consequence of these results.

As mentioned earlier, one drawback of the analysis is the crude manner in which the probability of event (66), that no terms outside of S are selected, is bounded. This gives rise to the $\sqrt{2/\pi}(k+1)/(\tau p^a)$ term in

the expression for $\tilde{p}_{err,k}$ (31), because of which a has to be greater than 1 when k is not negligible compared to p . In [14], a more careful analysis had been carried out for exact recovery with i.i.d. designs and ℓ_0 -sparse vectors. We believe that their analysis should carry over for the general case analyzed here, by noting that the random variables E_{ji} , for $i = 1, \dots, m+1$, defined in Lemma 4.2, has the same covariance structure as a normalized Brownian motion at times t_1, \dots, t_{m+1} , where $t_i = \|\tilde{R}_{i-1}\|^2$. This should improve the probability of the event (66) to something closer to $1/p^a$.

For random designs, we measure the performance after averaging over the distribution of X . As mentioned before, this can be contrasted to another method, as done in Candès and Plan [7] for the Lasso, in which a distribution is assigned to β and the performance is measured after averaging over this distribution. Although these two methods do not imply each other, it is interesting to compare the average performance using both methods. To be consistent with their notation, let's assume that the entries of X are scaled so that the columns have norm equal (or nearly equal) to one. Under a mild assumption on the incoherence, it is shown that for ℓ_0 -sparse vectors the support can be recovered, if

$$k = O(p/[\|X\|^2 \log p]), \quad (67)$$

where $\|X\|$ denotes the spectral norm of X . If X has i.i.d $N(0, 1/n)$ entries, then $\|X\| \approx \sqrt{p/n}$, so that the sparsity requirement (67) would translate to $k = O(n/\log p)$, which is of the same order as what we get here. However, the situation is different in the general case when the rows are i.i.d $N(0, \Sigma/n)$. Then X may be expressed as $\tilde{X}\Sigma^{1/2}$, where \tilde{X} has i.i.d $N(0, 1/n)$ entries. Consider the example where $\Sigma_{ii} = 1$ and $\Sigma_{ij} = c/k$, when $i \neq j$, with c appropriately chosen. In this case $\|X\| \approx c'p/\sqrt{nk}$. Consequently, (67) translates to assuming $n = \Omega(p \log p)$. Our results are better in this case, since we only require $\Omega(k \log p)$ observations even for such correlated designs.

An advantage of the work in [7] is its applicability to broad classes of deterministic designs. It is unclear at this stage whether such results also hold for the OMP.

A Tail bounds

A random variable Z is said to be sub-gaussian with mean 0 and scale $\sigma > 0$, if $\mathbb{E}e^{tZ} \leq e^{t^2\sigma^2/2}$ for each $t \in \mathbb{R}$.

Lemma A.1. *Let $W = (W_j : 1 \leq j \leq n)^T$, with each W_j sub-gaussian with mean 0 and scale $\sigma_j > 0$. Let $\sigma = \max_j \{\sigma_j\}$. The following hold.*

- (a) *Let $h \in \mathbb{R}^n$, with $\|h\| \leq 1$. If the entries of W are independent then $h^T W$ is sub-gaussian with mean 0 and scale σ .*

(b) Let $\rho = \sigma \sqrt{2(1+a) \log p}$ with $a > 0$. Then $P(\max_j |W_j| > \rho) \leq 2n/p^{1+a}$. Further, if the $W_j \sim N(0, \sigma^2)$ then this probability can be bounded by $\sqrt{2/\pi}(\sigma n)/(\rho p^{1+a})$.

Proof. For part (a), we need to show that $E \exp\{t h^\top W\} \leq \exp\{t^2 \sigma^2/2\}$. To see this, notice that $E \exp\{t h^\top W\} = E \exp\left\{t^2 \sum_{j=1}^n h_j^2 \sigma_j^2/2\right\}$, using independence of W_j 's. The claim is proved by noticing that $\sum_{j=1}^n h_j^2 \sigma_j^2/2 \leq \sigma^2$, using $\|h\| \leq 1$ and $\sigma_j \leq \sigma$.

For part (b), use a Chernoff bound, followed by optimizing the exponent to get that,

$$\mathbb{P}(|W_j| > \rho) \leq 2 \exp\left(-\frac{\rho^2}{2\sigma^2}\right).$$

If the W_j 's were normal, standard tail bounds [13] reveals that the above bound can be improved to $(2/(\sqrt{2\pi}\rho)) \exp\left(-\frac{\rho^2}{2\sigma^2}\right)$. Now use a union bound, along with the fact that $\exp\left(-\frac{\rho^2}{2\sigma^2}\right) = 1/p^{1+a}$, to prove the claim. \square

Next we give a simple lemma on chi-square tail bounds, which will be used repeatedly.

Lemma A.2. *Let W follow $N(0, I_n)$. Then*

$$\mathbb{P}(\|W\|/\sqrt{n} \geq 1 + \mu_n) \leq 1/p, \quad (68)$$

where $\mu_n = \sqrt{(2 \log p)/n}$.

Proof. Use the fact (see for example [11]) that for $h > 0$, one has

$$\mathbb{P}(\|W\|/\sqrt{n} \geq 1 + h) \leq e^{-nh^2/2}.$$

Substitute $h = \sqrt{(2 \log p)/n}$ to get the result. \square

B Proof of Lemma 3.1

For convenience, let $S = \{1, \dots, k\}$. Let H_j , $1 \leq j \leq k$ denote the columns of the H matrix. Assume that the algorithm runs for m steps and let R_1, \dots, R_{m-1} denote the associated residuals. Let $R_0 = Y$. Denote as $\hat{U}_{\mathcal{A}}$, the least square fit when U is regressed on $H_{\mathcal{A}}$. We also denote as $u(i) = S - d(i)$, which corresponds to the terms in S undetected after step i . We assume $u(0) = S$ and $\hat{U}_{d(0)} = 0$.

The following lemma is from Zhang [28].

Lemma B.1. (Zhang [28]) *For each i , with $0 \leq i < m$, if $|u(i)| > 0$, then*

$$\max_{j \in u(i)} \left| \frac{H_j^\top R_i}{\|H_j\|} \right| \geq \sqrt{\lambda_{\min}} \frac{\|\hat{U}_{d(i)} - \hat{U}_S\|}{\sqrt{|u(i)|}},$$

The results is a consequence of Lemmas 6 and 7 in Zhang [28, page 566]. Using his notation, in our case, $\lambda_{\min} = \rho(\bar{F})$, $R_i = Y - X\beta^{(k-1)}$, $\hat{U}_{d(i)} = X\beta^{(k-1)}$, $\hat{U}_S = X\beta_X(\bar{F}, y)$ and $u(i) = \bar{F} - F^{(k-1)}$.

Lemma B.2. *For each i , with $0 \leq i \leq m$, one has*

$$\|R_i\|/\sqrt{n} \leq \sqrt{\tilde{\lambda}_{\max}}(\|\varphi_{u(i)}\| + \sigma),$$

where $\tilde{\lambda}_{\max} = \max\{\lambda, \lambda_{\max}\}$.

Proof of B.2. Write $R_i = (I - \mathcal{P}_i)U$, where here \mathcal{P}_i is the projection matrix for column space of $H_{d(i)}$. Now $U = H_{d(i)}\varphi_{d(i)} + H_{u(i)}\varphi_{u(i)} + \epsilon$ and $(I - \mathcal{P}_i)H_{d(i)} = 0$. Correspondingly, $R_i = (I - \mathcal{P}_i)[H_{u(i)}\varphi_{u(i)} + \epsilon]$. Consequently, $\|R_i\| \leq \|H_{u(i)}\varphi_{u(i)}\| + \|\epsilon\|$, since $\|(I - \mathcal{P}_i)x\| \leq \|x\|$ for any $x \in \mathbb{R}^n$. The result immediately follows from using $\|H_{u(i)}\varphi_{u(i)}\|/\sqrt{n} \leq \sqrt{\lambda_{\max}}\|\varphi_{u(i)}\|$ and $\|\epsilon\|/(\sqrt{n}\sigma) \leq \sqrt{\lambda}$. This completes the proof of the lemma. \square

Now use the fact that $\|H_j\| \geq \sqrt{n}\sqrt{\lambda_{\min}}$, to get from Lemma B.1 that,

$$\max_{j \in u(i)} |H_j^T R_i| \geq \sqrt{\frac{n\rho_1}{|u(i)|}} \|\hat{U}_{d(i)} - \hat{U}_S\|,$$

where $\rho_1 = \lambda_{\min}^2$. Consequently, using Lemma B.2 and the above, one has that,

$$\max_{j \in u(i)} \left| H_j^T \frac{R_i}{\|R_i\|} \right| \geq \sqrt{\frac{n\rho_2}{|u(i)|}} \frac{\|\hat{U}_{d(i)} - \hat{U}_S\|/\sqrt{n}}{\|\varphi_{u(i)}\| + \sigma},$$

where $\rho_2 = \rho_1/\tilde{\lambda}_{\max}$. The algorithm continues as long as the left side of the above is at least τ_0 . Consequently, following the reasoning in [28], when the algorithm stops, one must have that either $|\hat{F}_2| = 0$ or the right side of the above, with $u(i)$ replaced by \hat{F}_2 , is at most τ_0 . Let's assume that $|\hat{F}_2| > 0$, since otherwise we would have correctly decoded all terms. Correspondingly, we have,

$$\|\hat{U}_{\hat{F}_2} - \hat{U}_S\|/\sqrt{n} \leq \tau_0 \sqrt{\frac{|\hat{F}_2|}{n\rho_2}} (\|\varphi_{\hat{F}_2}\| + \sigma) \quad (69)$$

when the algorithm stops. Now,

$$\|\varphi_{\hat{F}_2}\| \leq \sqrt{|\hat{F}_2|} \|\varphi - \hat{\varphi}_{ls}\|_\infty + \|\hat{\varphi}_{ls} - \hat{\varphi}\|. \quad (70)$$

To see this note that $\|\varphi_{\hat{F}_2}\|$ is bounded by the sum of $\|\varphi_{\hat{F}_2} - \hat{\varphi}_{ls, \hat{F}_2}\|$ and $\|\hat{\varphi}_{ls, \hat{F}_2}\|$, where $\hat{\varphi}_{ls, \hat{F}_2}$ is the sub-vector of $\hat{\varphi}_{ls}$ with indices in \hat{F}_2 . The first term in the bound is at most $\sqrt{|\hat{F}_2|} \|\varphi - \hat{\varphi}_{ls}\|_\infty$, whereas the second term can be bounded by $\|\hat{\varphi}_{ls} - \hat{\varphi}\|$, since $\hat{\varphi}_j$ is zero for all indices j in \hat{F}_2 . Now, use the fact that $\|\hat{\varphi}_{ls} - \varphi\|_\infty$ is bounded by $c_0\sigma\tau_0/\sqrt{n}$ along with the fact that $\|\hat{U}_{\hat{F}_2} - \hat{U}_S\|/\sqrt{n} \geq \sqrt{\lambda_{\min}}\|\hat{\varphi} - \hat{\varphi}_{ls}\|$, to get that from (69) and (70) that,

$$\|\varphi_{\hat{F}_2}\| \leq c_0\sigma\tau_0 \sqrt{\frac{|\hat{F}_2|}{n}} + \tau_0 \sqrt{r_1 \frac{|\hat{F}_2|}{n}} (\|\varphi_{\hat{F}_2}\| + \sigma) \quad (71)$$

when the algorithm stops. Here we use that $r_1 = 1/(\lambda_{\min}\rho_2)$. One gets from (71) that

$$\left(1 - \tau_0 \sqrt{\frac{r_1 |\hat{F}_2|}{n}}\right) \|\varphi_{\hat{F}_2}\| \leq \tilde{r}_2 \sigma \tau_0 \sqrt{|\hat{F}_2|/\sqrt{n}}, \quad (72)$$

where $\tilde{r}_2 = c_0 + \sqrt{r_1}$ and $r_1 = 1/\rho$. Using $|\hat{F}_2| \leq k$, the term $\tau_0 \sqrt{r_1 |\hat{F}_2|/n}$ appearing in the left side of the above can be bounded by $\tau_0 \sqrt{r_1 k/n}$. This leads us to (44), which completes the proof of part (a).

For part (b), notice that

$$\|\hat{\varphi} - \varphi\| \leq \sqrt{k} \|\hat{\varphi}_{ls} - \varphi\|_\infty + \|\hat{\varphi}_{ls} - \hat{\varphi}\|. \quad (73)$$

Now use,

$$\|\hat{\varphi}_{ls} - \hat{\varphi}\| \leq \tau_0 \sqrt{r_1 k/n} (\|\varphi_{\hat{F}_2}\| + \sigma)$$

along with,

$$\|\varphi_{\hat{F}_2}\| \leq \frac{\tilde{r}_2 \sigma \tau_0 \sqrt{k/n}}{\left(1 - \tau_0 \sqrt{r_1 k/n}\right)}, \quad (74)$$

to get, after rearranging, that,

$$\|\hat{\varphi}_{ls} - \hat{\varphi}\| \leq \sigma \tau_0 \sqrt{r_1 k/n} \frac{(c_0 \tau_0 \sqrt{k/n} + 1)}{1 - \tau_0 \sqrt{r_1 k/n}}.$$

Now use $\|\hat{\varphi}_{ls} - \varphi\|_\infty \leq \sigma c_0 \tau_0 \sqrt{k/n}$, along with $\tilde{r}_2 = c_0 + \sqrt{r_1}$, to get from (73) and the above that,

$$\|\hat{\varphi} - \varphi\| \leq \frac{\tilde{r}_2 \sigma \tau_0 \sqrt{k/n}}{\left(1 - \tau_0 \sqrt{r_1 k/n}\right)}.$$

This completes the proof of the lemma.

C Proof of Lemma 2.3

For a matrix $A \in \mathbb{R}^{n \times m}$, and $a = 1$ or ∞ , denote as $\|A\|_a = \sup_{v \neq 0} \|Av\|_a / \|v\|_a$. Recall that $\|A\|_1$ is the maximum of the ℓ_1 norms of the columns, whereas $\|A\|_\infty$ is the maximum of the ℓ_1 norms of the rows.

We first prove part (i). We use Cai and Wang [5, Lemma 2], to get that

$$1 - \gamma(k-1) \leq s_{\min} \leq s_{\max} \leq 1 + \gamma(k-1).$$

Now $\gamma \leq \omega_0/(2k)$, since $k \leq \bar{k}$, and hence, the left side of the above is at least $1 - \omega_0/2$ and the right side is at most $1 + \omega_0/2$. Further, use Tropp [23, Theorem 3.5], to get that

$$\|\Sigma_{SS}^{-1} \Sigma_{Sj}\|_1 \leq \frac{\gamma k}{1 - \gamma(k-1)}.$$

The right side of the above is at most ω_0 . Correspondingly, we may take ω as ω_0 .

We next prove part (ii). Use the fact that,

$$\|\Sigma_{SS}^{-1}\Sigma_{SS^c}\beta_{S^c}\|_\infty \leq \|\Sigma_{SS}^{-1}\|_\infty \|\Sigma_{SS^c}\beta_{S^c}\|_\infty. \quad (75)$$

Now as Σ_{SS}^{-1} is symmetric, $\|\Sigma_{SS}^{-1}\|_\infty = \|\Sigma_{SS}^{-1}\|_1$; the latter is at most $1/(1-\gamma(k-1))$ from [23, Theorem 3.5]. Further, $\|\Sigma_{SS^c}\beta_{S^c}\|_\infty \leq \gamma\|\beta_{S^c}\|_1$, which is at most $\sigma\gamma\eta\mu_n$. Correspondingly, from (75), one gets

$$\|\Sigma_{SS}^{-1}\Sigma_{SS^c}\beta_{S^c}\|_\infty \leq \sigma \frac{\gamma\bar{k}}{1-\gamma(k-1)} \bar{\eta}\mu_n. \quad (76)$$

The right of the above is at most $\sigma\omega_0\bar{\eta}\mu_n$, using the bound on γ . Further,

$$\|\Sigma_{S^c|S}\|_\infty \leq \|\Sigma_{S^cS^c}\beta_{S^c}\|_\infty + \|\Sigma_{S^cS}\Sigma_{SS}^{-1}\Sigma_{SS^c}\beta_{S^c}\|_\infty. \quad (77)$$

Now, $\|\Sigma_{S^cS^c}\beta_{S^c}\|_\infty \leq \|\beta_{S^c}\|_\infty + \|(\Sigma_{S^cS^c} - I)\beta_{S^c}\|_\infty$. Further, use $\|\beta_{S^c}\|_\infty \leq \sigma\nu\mu_n$ and $\|(\Sigma_{S^cS^c} - I)\beta_{S^c}\|_\infty \leq \gamma\|\beta_{S^c}\|_1$, the right side of which is at most $\sigma\gamma\eta\mu_n$. Also, the second term in (77) can be bounded as follows:

$$\|\Sigma_{S^cS}\Sigma_{SS}^{-1}\Sigma_{SS^c}\beta_{S^c}\|_\infty \leq \|\Sigma_{S^cS}\|_\infty \|\Sigma_{SS}^{-1}\Sigma_{SS^c}\beta_{S^c}\|_\infty.$$

The first term in the right side product is bounded by γk , whereas the second term, from (76), is bounded by $\sigma\omega_0\bar{\eta}\mu_n$. Correspondingly, one gets that

$$\|\Sigma_{S^c|S}\beta_{S^c}\|_\infty \leq \sigma\nu\mu_n + \sigma\gamma\eta\mu_n + \sigma\gamma\omega_0\eta\mu_n.$$

Further, using $\gamma\eta + \gamma\eta\omega_0 \leq 2\gamma\eta$, which is at most $\omega_0\bar{\eta}$, one gets the bound on $\|\Sigma_{S^c|S}\beta_{S^c}\|_\infty$.

For $k = 0$, one has $\|\Sigma_{S^cS^c}\beta_{S^c}\|_\infty \leq \nu + \omega_0\bar{\eta}$, which is at most $\nu + \omega_0\bar{\eta}$, from the bound derived above. This completes the proof of the lemma.

References

- [1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. The johnson-lindenstrauss lemma meets compressed sensing. *Constructive Approximation*, 2007.
- [2] A.R. Barron and A. Joseph. Sparse superposition codes: Fast and reliable at rates approaching capacity with gaussian noise. Technical report, Yale University, 2010.
- [3] A.R. Barron and A. Joseph. Least squares superposition coding of moderate dictionary size, reliable at rates up to channel capacity. *Submitted to IEEE Trans. Inform. Theory*, 2010.
- [4] A.R. Barron, A. Cohen, W. Dahmen, and R.A. DeVore. Approximation and learning by greedy algorithms. *Ann. Statist.*, 36(1):64–94, 2008.

- [5] T. Cai and L. Wang. Orthogonal matching pursuit for sparse signal recovery. Technical report, 2010.
- [6] E. Candes and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n . *Ann. Statist.*, 35(6):2313–2351, 2007.
- [7] E.J. Candès and Y. Plan. Near-ideal model selection by l_1 minimization. *Ann. Statist.*, 37(5A):2145–2177, 2009.
- [8] E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, 2005.
- [9] E.J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.
- [10] D.L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.
- [11] D.L. Donoho. For most large underdetermined systems of equations, the minimal l_1 -norm near-solution approximates the sparsest near-solution. *Communications on pure and applied mathematics*, 59(7):907–934, 2006.
- [12] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.
- [13] W. Feller. An introduction to probability theory and its applications. vol. i. 1950.
- [14] A.K. Fletcher and S. Rangan. Orthogonal matching pursuit: A brownian motion analysis. *Arxiv preprint. arXiv:1105.5853*, 2011.
- [15] C. Huang, G.H.L. Cheang, and A.R. Barron. Risk of penalized least squares, greedy selection and l_1 penalization for flexible function libraries. *Submitted to Ann. Statist.*, 2008.
- [16] L. Jones. A simple lemma for optimization in a hilbert space, with application to projection pursuit and neural net training. *Ann. Statist.*, 20:608–613, 1992.
- [17] W.S. Lee, P.L. Bartlett, and R.C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inform. Theory*, 42(6):2118–2132, 1996.
- [18] S. Mallat and S.M.Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Processing*, 41:3397–3415, 1993.
- [19] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

- [20] Y.C. Pati, R. Rezaifar, and PS Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conf. Rec. 27th Asilomar Conf. Sig., Sys. and Comput.*, pages 40–44. IEEE, 1993.
- [21] S.J. Szarek. Condition numbers of random matrices. *Journal of Complexity*, 7(2):131–149, 1991.
- [22] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, pages 267–288, 1996.
- [23] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.
- [24] J.A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 52(3):1030–1051, 2006.
- [25] J.A. Tropp and A.C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, 53(12):4655–4666, 2007.
- [26] M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso). *IEEE Trans. Inform. Theory*, 55(5):2183–2202, 2009.
- [27] C.H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.
- [28] T. Zhang. On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.*, 10:555–568, 2009.
- [29] T. Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *Ann. Statist.*, 37(5A):2109–2144, 2009.
- [30] P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.